SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 1–21

ON A MULTIPLE-SCALES ANALYSIS OF MULTILATERAL PHENOMENA IN SEMICONDUCTOR LASERS*

BENJAMIN P. COX[†] AND WARREN R. SMITH[‡]

Abstract. A mathematical model describing the coupling of electrical and optical effects in the active region of a realistic semiconductor laser medium is introduced. The weakly nonlinear analysis which follows gives rise to a leading-order problem describing three lateral modes. At the next order, the secularity conditions exhibit competition for photons and modal interaction. By making further assumptions, a partially lumped model is deduced which has no counterpart in the existing literature; this simplified system consists of one parabolic and six first-order hyperbolic partial differential equations. Predictions of this partially lumped model are compared with experimental observations.

Key words. semiconductor lasers, multilateral phenomena, asymptotics

AMS subject classifications. 35B25, 78A60

DOI. 10.1137/080720875

1. Introduction.

Two different approaches are followed by the laser community when modeling multimode effects in semiconductor lasers. The first approach describes M longitudinal modes by M rate equations (see, for example, [1], [2], and [11]). However, rate equations for the lateral modes have not been determined despite being required to model broad-area lasers. The second approach attempts to incorporate all possible effects (including sophisticated quantum mechanical models) into large numerical codes using the Maxwell–Bloch equations (see, for example, [7] and [10]). In this article, an alternative approach is proposed to determine simplified models based on a systematic asymptotic analysis of Maxwell's equations. Moreover, we derive traveling-wave rate equations for the lateral modes in a broad-area semiconductor laser.

A schematic cross-section of a typical semiconductor laser is shown in Figure 1. The current, assumed unidirectional in this paper, passes between the metal contact on the substrate and the heat sink. Electrons are injected into the active layer where they recombine with holes through both radiative and nonradiative mechanisms. During radiative recombination the energy released by an electron-hole pair appears in the form of a photon. This can happen through spontaneous emission, in which the photons are emitted in random directions, or stimulated emission, in which recombination is initiated by an existing photon. In the latter case the emitted photon matches the original photon in wavelength, phase, and direction. As long as the end faces of the semiconductor possess a suitable reflectivity and the current exceeds a given threshold value, the semiconductor is excited through stimulated emission into laser operation.

The purpose of this paper is to investigate the multilateral effects in semiconductor lasers by extending the weakly nonlinear analysis presented in [12]. The semiconductor laser is split into two regions—the active layer, in which the electrical-optical effects

*Received by the editors June 20, 2005; accepted for publication (in revised form) April 25, 2008; published electronically July 25, 2008.
http://www.siam.org/journals/siap/69-1/63406.html

[†]Tessella, Elopak House, Rutherford Close, Meadway Technology Park, Stevenage, Hertfordshire, SG1 2EF, United Kingdom (Ben.Cox@tessella.com).

[‡]School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom (smithwar@for.mat.bham.ac.uk).

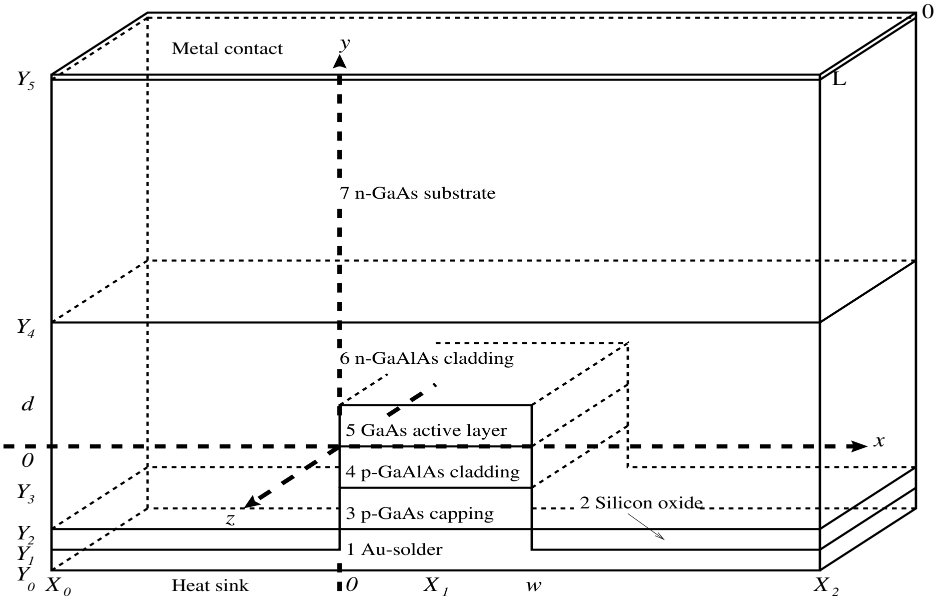


FIG. 1. *Typical layered structure of a double-heterostructure semiconductor laser. The lateral direction is denoted by x, the transverse direction by y, and the longitudinal (or axial) direction by z. The battery potential is applied in the transverse direction, with mirrors located at $z = 0$ and $z = L$.*

take place and lasing occurs, and the surrounding material. This work is concerned with the first of these regions. The analysis in [12], which assumes one dominant lateral mode, is extended to the case of three such modes. As multilateral mode operation is generally undesirable, we also investigate a bifurcation condition to quantify the onset of multilateral behavior.

We take Maxwell's equations as the starting point and add models for the polarization and current density. We assume that the whole device is maintained at a constant ambient temperature and that the gain takes place at a single frequency associated with the band gap of the active region. Each of the modes must satisfy the periodicity condition $\lambda_r = 2nL/q_r$, where n and q_r are positive integers, λ_r is the wavelength of the rth lateral mode, and L is the longitudinal cavity length. We assume that each lasing mode is confined to the active layer, noting that in certain devices, such as quantum well lasers, this is true of only a tenth of the lasing mode—the analysis in this paper will not be valid in such circumstances. We are primarily concerned with the steady-state behavior which is insensitive to spontaneous emission. Hence the contribution due to this process is omitted in the derivation of the models below.

Based on the above assumptions, the mathematical model is summarized in section 2. The problem is nondimensionalized in section 3, where the key small parameters are identified. Section 4 deals with a multiple-scale asymptotic analysis, whereby the governing equations are reduced to one parabolic partial differential equation and twelve first-order wave equations. In section 5, further simplifications lead to a time-dependent partially lumped model whose bifurcations are investigated. Numerical solutions to the steady-state system are presented in section 6. We split the results into two cases, prescribing a different lateral current density profile in each case, and compare the physical effects predicted by the model with experimental observations. The final section briefly draws some conclusions.

2. Theoretical background. The model formulation in the active region will now be outlined. We write down Maxwell's equations for a semiconductor medium in the form

$$\boldsymbol{\nabla}\cdot\boldsymbol{D} = \rho, \tag{1}$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{2}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{B} = 0, \tag{3}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{J}, \tag{4}$$

where $\boldsymbol{E}$ is the electric field, $\boldsymbol{D}$ is the electric displacement, $\boldsymbol{H}$ is the magnetic field, $\boldsymbol{B}$ is the magnetic induction, $\boldsymbol{J}$ is the current density, ρ is the charge density, t is time, and the differential operator $\boldsymbol{\nabla} = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$, where x, y, and z are indicated in Figure 1. The constitutive equations are given by $\boldsymbol{D} = \epsilon_0\boldsymbol{E} + \boldsymbol{P}$ and $\boldsymbol{B} = \mu_0\boldsymbol{H}$, where ϵ_0 is the permittivity of free space, $\boldsymbol{P}$ is the polarization, and μ_0 is the permeability of free space. The speed of light in a vacuum is given by $c_0 = 1/\sqrt{\epsilon_0\mu_0}$.

We write the polarization as a sum,

$$\boldsymbol{P} = (\epsilon - \epsilon_0)\boldsymbol{E} + \boldsymbol{P}^{(R)}, \tag{5}$$

where the permittivity ϵ is assumed to be independent of changes in the carrier concentration (and self-focusing is neglected as a result). The time-periodic component $\boldsymbol{P}^{(R)}$ represents resonant coupling, which occurs when the frequency of the electric field closely matches a natural frequency of the semiconductor. As in [12] we adopt a simple model for this resonant interaction,

$$\frac{\partial^2\boldsymbol{P}^{(R)}}{\partial t^2} = -\frac{\partial}{\partial t}(\epsilon a(\min(n,p) - n_t)\boldsymbol{E}), \tag{6}$$

where n and p are the electron and hole concentrations, respectively, $\epsilon a n_t$ represents the absorption of photons, $-\epsilon a \min(n,p)$ is the stimulated gain, a is the linear gain rate, and n_t is the electron density at transparency (that is, the electron concentration at which the stimulated gain and absorption balance). Equation (6) is consistent with the corresponding constitutive equations adopted elsewhere (see [1]) and can be regarded as an approximation to the more sophisticated theories of polarization (see, for example, [3] and [10]).

The charge density and current density are split up as follows:

$$\rho = e(N + p - n), \qquad \boldsymbol{J} = \boldsymbol{J}_{\boldsymbol{n}} + \boldsymbol{J}_{\boldsymbol{p}}, \tag{7}$$

where e is the charge on an electron, N is the net impurity density in the active region, and $\boldsymbol{J}_{\boldsymbol{n}}$ and $\boldsymbol{J}_{\boldsymbol{p}}$ are the components of the current density carried by electrons and holes, respectively. If we let D_n and D_p represent the active region diffusivities of electrons and holes, respectively, and use μ_n and μ_p to represent their respective mobilities, then the components of the current density are written in the form (see [12])

$$\boldsymbol{J}_{\boldsymbol{n}} = e\,(D_n\boldsymbol{\nabla} n + \mu_n n\boldsymbol{E}), \qquad \boldsymbol{J}_{\boldsymbol{p}} = e\,(-D_p\boldsymbol{\nabla} p + \mu_p p\boldsymbol{E}). \tag{8}$$

Defining k to be Boltzmann's constant and T to be the temperature of the device (assumed constant here), the Einstein relations between the diffusivities and mobilities are $D_n = U_T \mu_n$ and $D_p = U_T \mu_p$, where the thermal voltage is given by $U_T = kT/e$. The continuity equations for electrons and holes are derived by taking the divergence of (4) and can be written in the form

$$e\frac{\partial n}{\partial t} = e(G - R) + \boldsymbol{\nabla}\boldsymbol{\cdot}\boldsymbol{J_n}, \qquad e\frac{\partial p}{\partial t} = e(G - R) - \boldsymbol{\nabla}\boldsymbol{\cdot}\boldsymbol{J_p}. \tag{9}$$

The net recombination rate of electron-hole pairs, $R - G$, is given by

$$\begin{aligned} R - G = {} & \frac{(np - n_i^2)}{A_p(n + n_i) + A_n(p + n_i)} + B(np - n_i^2) \\ & + (C_n n + C_p p)(np - n_i^2) - \frac{1}{E_g}\boldsymbol{E}\cdot\frac{\partial \boldsymbol{P}^{(R)}}{\partial t}. \end{aligned} \tag{10}$$

The terms on the right-hand side of (10) correspond respectively to Shockley–Read–Hall recombination (with reciprocal rate constants A_n and A_p), radiative recombination (with rate constant B), Auger recombination (with rate constants C_n and C_p), and recombination due to stimulated emission. We denote the intrinsic carrier density by n_i, while the constant E_g corresponds to the band-gap energy in the active region. We assume that the laser operates at a single angular frequency $\omega = 2\pi E_g/h$, where h represents Planck's constant.

Equations (1)–(4) can be simplified by the introduction of scalar and vector potentials. We satisfy (2) and (3) by writing $\boldsymbol{E} = -\boldsymbol{\nabla}\phi - \partial \boldsymbol{A}/\partial t$ and $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ in which ϕ is the scalar potential and $\boldsymbol{A}$ is the vector potential. We specify $\boldsymbol{A}$ uniquely via the gauge condition (see [14])

$$\boldsymbol{\nabla}\boldsymbol{\cdot}\boldsymbol{A} + \mu_0 \epsilon \frac{\partial \phi}{\partial t} = 0 \tag{11}$$

so that ϕ and $\boldsymbol{A}$ satisfy

$$\begin{aligned} \boldsymbol{\nabla}^2 \phi - \mu_0 \epsilon \frac{\partial^2 \phi}{\partial t^2} &= \frac{1}{\epsilon}\Big(\boldsymbol{\nabla}\boldsymbol{\cdot}\boldsymbol{P}^{(R)} - e(N + p - n)\Big), \\ \boldsymbol{\nabla}^2 \boldsymbol{A} - \mu_0 \epsilon \frac{\partial^2 \boldsymbol{A}}{\partial t^2} &= -\mu_0 \frac{\partial \boldsymbol{P}^{(R)}}{\partial t} - \mu_0 \boldsymbol{J}. \end{aligned} \tag{12}$$

The mathematical model comprises (6), (9), and (11)–(12) along with appropriate boundary conditions (see [12]).

3. The dimensionless mathematical model. In the nondimensionalization which follows, the quantities ϕ_e and A_e (defined in [12]) represent typical magnitudes of the scalar and vector potentials, respectively. We define n_e to be a representative value of the electron concentration (defined in [12]). Let w be the lateral width of the active region, d be the transverse width of the active region, λ be a typical wavelength of the lasing radiation, and $\tau_e = \lambda/c_0$ be the time-scale. We transform to dimensionless variables via $n = n_e \hat{n}$, $p = n_e \hat{p}$, $\phi = \phi_e \hat{\phi}$, $\boldsymbol{A} = A_e \hat{\boldsymbol{A}}$, $\boldsymbol{P}^{(R)} = n_e \lambda e \delta \nu \hat{\boldsymbol{P}}^{(R)}$, $x = w\hat{x}$, $y = d\hat{y}$, $z = \lambda \hat{z}$, and $t = \tau_e \hat{t}$, where the dimensionless constants δ and ν are defined in Table 1.

The model consists of continuity equations for the electron and hole densities $\hat{n}$ and $\hat{p}$, two second-order wave equations modeling the electric field components $\hat{\phi}$ and

Table 1

Dimensionless parameters for a typical GaAs laser diode (note that some values are reliable, while others are best estimates).

Symbol	Definition	Typical value
δ	λ/L	4×10^{-3}
ν	$L\sqrt{\epsilon\mu_0}/A_n$	2×10^{-4}
A_x	λ/w	0.1
A_y	λ/d	2
$\delta^{-1}\nu^{-1}\mathcal{A}_n$	A_n/τ_e	3×10^{6}
$\delta^{-1}\nu^{-1}\mathcal{A}_p$	A_p/τ_e	3×10^{6}
$\delta\nu\mathcal{B}$	$Bn_e\tau_e$	6×10^{-7}
$\delta\nu\mathcal{C}_n$	$C_n n_e^2\tau_e$	4×10^{-7}
$\delta\nu\mathcal{C}_p$	$C_p n_e^2\tau_e$	4×10^{-7}
$\delta\nu\mathcal{D}_n$	$D_n\tau_e/\lambda^2$	3×10^{-8}
$\delta\nu\mathcal{D}_p$	$D_p\tau_e/\lambda^2$	2×10^{-9}
ϵ_R		10
$\mathcal{F}$	$A_e e\lambda/\tau_e E_g$	2
$\nu\mathcal{G}$	$\epsilon_0\phi_e/e\lambda^2 n_e$	6×10^{-5}
$\delta\nu\mathcal{H}$	$a\epsilon_0 A_e/e\lambda$	3×10^{-7}
$\nu^{-1}\mathcal{L}$	$n_e\mu_0 e\lambda^3/A_e\tau_e$	2×10^{4}
$\mathcal{N}$	n_i/n_e	3×10^{-11}
n^*	n_t/n_e	0.4
N^*	N/n_e	1×10^{-2}
$\delta\nu\mathcal{V}_n$	$\mu_n A_e/\lambda$	2×10^{-6}
$\delta\nu\mathcal{V}_p$	$\mu_p A_e/\lambda$	2×10^{-7}
$R^{(1)}$, $R^{(2)}$		0.3

$\hat{\boldsymbol{A}}$, the gauge condition, and an equation for the polarization component $\hat{\boldsymbol{P}}^{(R)}$:

$$\begin{aligned}\frac{\partial\hat{n}}{\partial\hat{t}} &= -\frac{\delta\nu(\hat{n}\hat{p}-\mathcal{N}^2)}{\mathcal{A}_p(\hat{n}+\mathcal{N})+\mathcal{A}_n(\hat{p}+\mathcal{N})} - \delta\nu\mathcal{B}(\hat{n}\hat{p}-\mathcal{N}^2) - \delta\nu(\mathcal{C}_n\hat{n}+\mathcal{C}_p\hat{p})(\hat{n}\hat{p}-\mathcal{N}^2)\\ &\quad - \delta\nu\mathcal{F}\frac{\partial\hat{\boldsymbol{P}}^{(R)}}{\partial\hat{t}}\cdot\left(\hat{\boldsymbol{\nabla}}\hat{\phi}+\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right) + \delta\nu\hat{\boldsymbol{\nabla}}\cdot\left(\mathcal{D}_n\hat{\boldsymbol{\nabla}}\hat{n} - \hat{n}\mathcal{V}_n\left\{\hat{\boldsymbol{\nabla}}\hat{\phi}+\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right\}\right),\end{aligned}\tag{13}$$

$$\begin{aligned}\frac{\partial\hat{p}}{\partial\hat{t}} &= -\frac{\delta\nu(\hat{n}\hat{p}-\mathcal{N}^2)}{\mathcal{A}_p(\hat{n}+\mathcal{N})+\mathcal{A}_n(\hat{p}+\mathcal{N})} - \delta\nu\mathcal{B}(\hat{n}\hat{p}-\mathcal{N}^2) - \delta\nu(\mathcal{C}_n\hat{n}+\mathcal{C}_p\hat{p})(\hat{n}\hat{p}-\mathcal{N}^2)\\ &\quad - \delta\nu\mathcal{F}\frac{\partial\hat{\boldsymbol{P}}^{(R)}}{\partial\hat{t}}\cdot\left(\hat{\boldsymbol{\nabla}}\hat{\phi}+\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right) + \delta\nu\hat{\boldsymbol{\nabla}}\cdot\left(\mathcal{D}_p\hat{\boldsymbol{\nabla}}\hat{p} + \hat{p}\mathcal{V}_p\left\{\hat{\boldsymbol{\nabla}}\hat{\phi}+\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right\}\right),\end{aligned}\tag{14}$$

$$\nu\epsilon_R\mathcal{G}\left(\hat{\boldsymbol{\nabla}}^2\hat{\phi}-\epsilon_R\frac{\partial^2\hat{\phi}}{\partial\hat{t}^2}\right) = \delta\nu\hat{\boldsymbol{\nabla}}\cdot\hat{\boldsymbol{P}}^{(R)} - (N^*+\hat{p}-\hat{n}),\tag{15}$$

$$\begin{aligned}\hat{\boldsymbol{\nabla}}^2\hat{\boldsymbol{A}} - \epsilon_R\frac{\partial^2\hat{\boldsymbol{A}}}{\partial\hat{t}^2} &= -\delta\mathcal{L}\left(\frac{\partial\hat{\boldsymbol{P}}^{(R)}}{\partial\hat{t}} + \mathcal{D}_n\hat{\boldsymbol{\nabla}}\hat{n} - \mathcal{D}_p\hat{\boldsymbol{\nabla}}\hat{p} - \mathcal{V}_n\hat{n}\hat{\boldsymbol{\nabla}}\hat{\phi} - \mathcal{V}_p\hat{p}\hat{\boldsymbol{\nabla}}\hat{\phi}\right.\\ &\qquad\left. - \mathcal{V}_n\hat{n}\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}} - \mathcal{V}_p\hat{p}\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right),\end{aligned}\tag{16}$$

$$\hat{\boldsymbol{\nabla}}\cdot\hat{\boldsymbol{A}} + \epsilon_R\frac{\partial\hat{\phi}}{\partial\hat{t}} = 0,\tag{17}$$

$$\frac{\partial^2\hat{\boldsymbol{P}}^{(R)}}{\partial\hat{t}^2} = \frac{\partial}{\partial\hat{t}}\left(\epsilon_R\hat{g}(\hat{n},\hat{p})\left(\hat{\boldsymbol{\nabla}}\hat{\phi}+\frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right)\right),\tag{18}$$

where the linear gain term in (18) is given by $\hat{g}(\hat{n},\hat{p}) = \mathcal{H}(\min(\hat{n},\hat{p}) - n^*)$. We define $A_x = \lambda/w$ and $A_y = \lambda/d$ whereby $\hat{\boldsymbol{\nabla}} = (A_x\partial/\partial\hat{x}, A_y\partial/\partial\hat{y}, \partial/\partial\hat{z})$. The relative permittivity, approximated by a constant in this weakly nonlinear analysis, is typically of $O(1)$ and $\epsilon_R = \epsilon/\epsilon_0$ (see Table 1). The dimensionless constants $\delta^{-1}\nu^{-1}\mathcal{A}_n$, $\delta^{-1}\nu^{-1}\mathcal{A}_p$, $\delta\nu\mathcal{B}$, $\delta\nu\mathcal{C}_n$, $\delta\nu\mathcal{C}_p$, $\delta\nu\mathcal{D}_n$, $\delta\nu\mathcal{D}_p$, $\delta\nu\mathcal{V}_n$, $\delta\nu\mathcal{V}_p$, $\mathcal{F}$, $\nu\mathcal{G}$, $\delta\nu\mathcal{H}$, $\nu^{-1}\mathcal{L}$, $\mathcal{N}$, n^*, and N^* are defined in Table 1; the constraints $\nu \ll \delta \ll 1$, $\mathcal{N} \ll 1$, and $N^* \ll 1$ typically hold in practice. The problem is regularly perturbed in $\mathcal{N}$ and N^*, and thus the terms that are multiplied by these are henceforth neglected. The former is the ratio of the intrinsic carrier density to a typical free carrier concentration. The latter represents the (low) doping concentration in the active region relative to the free carrier concentration. The problem is singularly perturbed in δ and ν. The first of these, δ, corresponds to the ratio of wavelength to cavity length; the second, ν, is the ratio of L/c to the time-scale on which the electron concentration varies (c is the speed of an electromagnetic wave in the active region medium). In the next section we will pursue a multiple-scale asymptotic analysis based around these key small parameters. Further discussion of the physical mechanisms and details that appear in the derivation of (13)–(18) may be found in [12].

We now outline the boundary conditions that accompany equations (13)–(18). In the following $\hat{\boldsymbol{n}}$ denotes the unit normal to the boundary of the lasing region. The normal components of the electron and hole current densities are assumed to be zero at the lateral boundaries, in which case

$$\left(\mathcal{D}_n\hat{\boldsymbol{\nabla}}\hat{n} - \mathcal{V}_n\hat{n}\left(\hat{\boldsymbol{\nabla}}\hat{\phi} + \frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right)\right)\cdot\hat{\boldsymbol{n}} = 0 \quad \text{at } \hat{x} = 0, 1, \tag{19}$$

together with a similar expression for holes. We prescribe the normal components of the electron and hole current densities at the transverse boundaries, that is,

$$\left(\mathcal{D}_n\hat{\boldsymbol{\nabla}}\hat{n} - \mathcal{V}_n\hat{n}\left(\hat{\boldsymbol{\nabla}}\hat{\phi} + \frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right)\right)\cdot\hat{\boldsymbol{n}} = \begin{cases} G_1(\hat{x},\hat{z},\hat{t}) & \text{at } \hat{y} = 0, \\ G_2(\hat{x},\hat{z},\hat{t}) & \text{at } \hat{y} = 1, \end{cases} \tag{20}$$

together with a similar expression for holes. At each mirror the normal fluxes of electrons and holes, denoted by $\partial\hat{n}/\partial\hat{z}$ and $\partial\hat{p}/\partial\hat{z}$, respectively, are assumed to be zero. The tangential component of the electric field and the normal component of the magnetic induction are assumed to be zero at the interfaces between lasing and nonlasing regions, and so

$$\hat{\boldsymbol{n}} \times \left(-\hat{\boldsymbol{\nabla}}\hat{\phi} - \frac{\partial\hat{\boldsymbol{A}}}{\partial\hat{t}}\right) = \boldsymbol{0} \quad \text{at } \ \hat{x} = 0, 1 \ \text{ and at } \ \hat{y} = 0, 1, \tag{21}$$

$$(\hat{\boldsymbol{\nabla}} \times \hat{\boldsymbol{A}})\cdot\hat{\boldsymbol{n}} = 0 \quad \text{at } \ \hat{x} = 0, 1 \ \text{ and at } \ \hat{y} = 0, 1. \tag{22}$$

In the models derived below it is possible to employ boundary conditions at the mirrors which approximate the reflected wave as a fraction of the incident wave (see (37)–(38)); this avoids the additional complication of modeling the transmitted electromagnetic wave.

4. Asymptotic analysis.

4.1. Periodicity. We require an important assumption concerning the multilateral mode case. The conjecture is that there exists an integer number of round trips on which the system is periodic (see section 1). Thus, when three lateral modes are operating within the laser's active region, we have

$$q_1\left(\frac{2\pi}{\hat{k}_1}\right) = q_2\left(\frac{2\pi}{\hat{k}_2}\right) = q_3\left(\frac{2\pi}{\hat{k}_3}\right),$$

where q_1, q_2, and q_3 are positive integers, while $\hat{k}_1$, $\hat{k}_2$, and $\hat{k}_3$ are the dimensionless wave numbers associated with the three modes. The corresponding dimensionless wavelengths are given by $\hat{\lambda}_r = 2\pi/\hat{k}_r$, where $\hat{\lambda}_r = \lambda_r/\lambda$. We will use r to differentiate between these lateral modes throughout the remainder of the paper.

4.2. Multiple scales. A multiple-scale asymptotic expansion is appropriate since there are two relevant longitudinal length-scales in the laser. These are wavelength ($\lambda \sim 10^{-6}$m) and cavity length ($L \sim 250 \times 10^{-6}$m). There are three physical time-scales to identify: the time-scale for carrier variations ($\sim 10^{-8}$s), that for an electromagnetic wave to traverse the longitudinal cavity length ($\sim 10^{-12}$s), and the reciprocal of the frequency ($\sim 10^{-15}$s). We shall employ a total of seven independent variables. Four of these describe space: the lateral length-scale $\hat{x}$, the transverse length-scale $\hat{y}$, the smaller axial length-scale $\hat{z}$, and the longer axial length-scale Z, where $Z = \delta\hat{z}$; the third variable corresponds to wavelength and the fourth to cavity length. The other three variables describe time: the shortest time-scale $\hat{t} = O(1)$, corresponding to the time-scale for a wave to travel a wavelength; the intermediate time-scale $T = \delta\hat{t} = O(1)$, the time-scale for traversing the cavity; and the longest time-scale $\tau = \delta\nu\hat{t} = O(1)$, that for carrier variations. At leading order the solution will be periodic in $\hat{t}$ with period $2\pi/\hat{\omega}$ ($\hat{\omega} = \omega\tau_e$ is the (known) dimensionless angular frequency) and in T with period $2q_2q_3\hat{k}_1/\hat{\omega}$. We note from Table 1 that $\nu \ll \delta$, and so we introduce expansions of the form $\hat{n} \sim n_0 + \delta n_1 + \nu n_2 + \delta^2 n_3 + \delta\nu n_4$, $\hat{p} \sim p_0 + \delta p_1$, $\hat{\boldsymbol{A}} \sim \boldsymbol{A_0} + \delta\boldsymbol{A_1}$, $\hat{\boldsymbol{P}}^{(R)} \sim \boldsymbol{P_0}$, and $\hat{\phi} \sim \phi_0 + \delta\phi_1$ as $\delta, \nu \to 0$. We seek solutions of the form $\phi_0 = \phi_0(\hat{x}, \hat{y}, \hat{z}, Z, \tau)$ and $\phi_1 = \phi_1(\hat{x}, \hat{y}, \hat{z}, Z, T, \tau)$ and consider only the y-component of the lasing mode to be nonzero at leading order, the dominance of the transverse electric mode being observed experimentally [16]. Hence we write $\boldsymbol{A_0} = (0, A_0^{(2)}, 0)$ and $\boldsymbol{A_1} = (0, A_1^{(2)}, 0)$.

Evaluating (13)–(14) at $O(1)$ and $O(\delta)$ implies that $n_0 = n_0(\hat{x}, \hat{y}, \hat{z}, Z, \tau)$ and $p_0 = p_0(\hat{x}, \hat{y}, \hat{z}, Z, \tau)$ with $n_0 = p_0$ given, at leading order, by (15). Since periodicity in $\hat{t}$ requires

$$\int_{\hat{t}=0}^{2\pi/\hat{\omega}} \frac{\partial n_4}{\partial \hat{t}}\, d\hat{t} = 0,$$

from (13) we have

$$\frac{\partial n_0}{\partial \tau} + \frac{\partial n_2}{\partial T} = -\frac{n_0}{\mathcal{A}_n + \mathcal{A}_p} - \mathcal{B}n_0^2 - (\mathcal{C}_n + \mathcal{C}_p)n_0^3 + \mathcal{D}_n\hat{\boldsymbol{\nabla}}^2 n_0 - \mathcal{V}_n\hat{\boldsymbol{\nabla}}\cdot(n_0\hat{\boldsymbol{\nabla}}\phi_0)$$
$$- \epsilon_R\mathcal{F}g(n_0)\,\frac{\hat{\omega}}{2\pi}\int_{\hat{t}=0}^{2\pi/\hat{\omega}}\left(\frac{\partial A_0^{(2)}}{\partial \hat{t}}\right)^2 d\hat{t},$$

where $g(n_0) = \mathcal{H}(n_0 - n^*)$ and we have assumed that $\partial\boldsymbol{P_0}/\partial\hat{t}$ is time-harmonic [12]. Periodicity in T demands that

$$\int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}} \frac{\partial n_2}{\partial T}\, dT = 0,$$

and so we obtain

$$\begin{aligned}\frac{\partial n_0}{\partial \tau} &= -\frac{n_0}{\mathcal{A}_n + \mathcal{A}_p} - \mathcal{B} n_0^2 - (\mathcal{C}_n + \mathcal{C}_p) n_0^3 + \mathcal{D}_n \hat{\boldsymbol{\nabla}}^2 n_0 - \mathcal{V}_n \hat{\boldsymbol{\nabla}} \boldsymbol{\cdot} (n_0 \hat{\boldsymbol{\nabla}} \phi_0) \\ &\quad - \epsilon_R \mathcal{F} g(n_0)\, \frac{\hat{\omega}^2}{4\pi q_2 q_3 \hat{k}_1} \int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}} \int_{\hat{t}=0}^{2\pi/\hat{\omega}} \left(\frac{\partial A_0^{(2)}}{\partial \hat{t}}\right)^2 d\hat{t}\, dT.\end{aligned} \tag{23}$$

A similar equation for $\partial p_0/\partial \tau$ is derived from (14). Subtracting this from (23) yields

$$\hat{\boldsymbol{\nabla}} \boldsymbol{\cdot} (n_0 \hat{\boldsymbol{\nabla}} \phi_0) = \left(\frac{\mathcal{D}_n - \mathcal{D}_p}{\mathcal{V}_n + \mathcal{V}_p}\right) \hat{\boldsymbol{\nabla}}^2 n_0,$$

which allows us to eliminate ϕ_0 from (23):

$$\begin{aligned}\frac{\partial n_0}{\partial \tau} &= -\frac{n_0}{\mathcal{A}_n + \mathcal{A}_p} - \mathcal{B} n_0^2 - (\mathcal{C}_n + \mathcal{C}_p) n_0^3 + \frac{\mathcal{D}_n \mathcal{V}_p + \mathcal{D}_p \mathcal{V}_n}{\mathcal{V}_n + \mathcal{V}_p} \hat{\boldsymbol{\nabla}}^2 n_0 \\ &\quad - \epsilon_R \mathcal{F} g(n_0)\, \frac{\hat{\omega}^2}{4\pi q_2 q_3 \hat{k}_1} \int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}} \int_{\hat{t}=0}^{2\pi/\hat{\omega}} \left(\frac{\partial A_0^{(2)}}{\partial \hat{t}}\right)^2 d\hat{t}\, dT.\end{aligned} \tag{24}$$

The first, second, third, and fifth terms on the right-hand side of (24) correspond to the electron-hole recombination mechanisms described in (10). The local reduction of electrons and holes due to spatial-hole burning (the integral term on the right-hand side of (24)) thus competes with the smoothing action of diffusion (fourth term on the right-hand side of (24)). The latter is the source of the current density term which will be introduced in section 5.

The following leading-order problem for the electric field is posed by (16)–(17):

$$\Box\, A_0^{(2)} = 0, \tag{25}$$

where the d'Alembertian operator is defined by

$$\Box = \epsilon_R \frac{\partial^2}{\partial \hat{t}^2} - A_x^2 \frac{\partial^2}{\partial \hat{x}^2} - \frac{\partial^2}{\partial \hat{z}^2}.$$

It follows from (21) that

$$A_0^{(2)} = 0 \quad \text{on} \quad \hat{x} = 0, 1. \tag{26}$$

We apply the method of separation of variables to (25)–(26). The solution is expressed in the form of an infinite sum in which each term corresponds to a lateral mode contained by the waveguide. For the remainder of this paper we consider three such modes which are distinguished using $r = 1, 2, 3$; their respective lateral profiles are $\sin(\pi\hat{x})$, $\sin(2\pi\hat{x})$, and $\sin(3\pi\hat{x})$. In this case

$$\begin{aligned}A_0^{(2)} = \sum_{r=1}^{3} \Big(& A_r^+ \cos(\hat{\omega}\hat{t} - \hat{k}_r\hat{z}) + B_r^+ \sin(\hat{\omega}\hat{t} - \hat{k}_r\hat{z}) \\ & + A_r^- \cos(\hat{\omega}\hat{t} + \hat{k}_r\hat{z}) + B_r^- \sin(\hat{\omega}\hat{t} + \hat{k}_r\hat{z})\Big) \sin(r\pi\hat{x}).\end{aligned} \tag{27}$$

The dispersion relation associated with (25) takes the form $\hat{k}_r^2 = \hat{\omega}^2\epsilon_R - (r\pi A_x)^2$. Evaluating (16)–(17) at $O(\delta)$ implies

$$\begin{aligned}\Box\, A_1^{(2)} &= 2\,\frac{\partial^2 A_0^{(2)}}{\partial\hat{z}\partial Z} - 2\epsilon_R\frac{\partial^2 A_0^{(2)}}{\partial\hat{t}\partial T} + \epsilon_R\mathcal{L}g(n_0)\frac{\partial A_0^{(2)}}{\partial\hat{t}} \\ &\quad + \mathcal{L}\left((\mathcal{D}_n - \mathcal{D}_p)A_y\frac{\partial n_0}{\partial\hat{y}} - (\mathcal{V}_n + \mathcal{V}_p)n_0A_y\frac{\partial\phi_0}{\partial\hat{y}} - (\mathcal{V}_n + \mathcal{V}_p)n_0\frac{\partial A_0^{(2)}}{\partial\hat{t}}\right),\end{aligned} \tag{28}$$

from which we deduce twelve secularity conditions governing the twelve unknown amplitude envelopes $A_r^{\pm}(Z,T,\tau)$ and $B_r^{\pm}(Z,T,\tau)$. Each condition takes the form of a first-order wave equation and can been verified using two different approaches. The first technique involves applying the Fredholm alternative to (28), and the second makes use of several trigonometric substitutions in order to directly eliminate the secular terms. We present the equations for A_1^+, A_2^+, and A_3^+ below; the equations for B_1^+, B_2^+, and B_3^+ are given in Appendix A. By symmetry in the direction of propagation, the equations for A_1^-, A_2^-, A_3^-, B_1^-, B_2^-, and B_3^- may be derived from (29)–(31) and (48)–(50) by replacing the quantities $\hat{k}_r$, A_r^+, B_r^+, A_r^-, and B_r^- with $-\hat{k}_r$, A_r^-, B_r^-, A_r^+, and B_r^+, respectively. We have

$$\begin{aligned}&\frac{2\pi q_1}{\hat{k}_1}\left(\epsilon_R\hat{\omega}\frac{\partial A_1^+}{\partial T} + \hat{k}_1\frac{\partial A_1^+}{\partial Z}\right) = A_1^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1-\alpha_2\big)\,d\hat{z} \\ &+ B_1^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1-\alpha_2\big)\sin\big(2\hat{k}_1\hat{z}\big)\,d\hat{z} + A_1^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1-\alpha_2\big)\cos\big(2\hat{k}_1\hat{z}\big)\,d\hat{z} \\ &+ B_3^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z} + A_3^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z} \\ &+ B_3^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z} + A_3^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z},\end{aligned} \tag{29}$$

$$\begin{aligned}&\frac{2\pi q_2}{\hat{k}_2}\left(\epsilon_R\hat{\omega}\frac{\partial A_2^+}{\partial T} + \hat{k}_2\frac{\partial A_2^+}{\partial Z}\right) = A_2^+\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\,d\hat{z} \\ &+ B_2^-\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\sin\big(2\hat{k}_2\hat{z}\big)\,d\hat{z} + A_2^-\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\cos\big(2\hat{k}_2\hat{z}\big)\,d\hat{z},\end{aligned} \tag{30}$$

$$\begin{aligned}&\frac{2\pi q_3}{\hat{k}_3}\left(\epsilon_R\hat{\omega}\frac{\partial A_3^+}{\partial T} + \hat{k}_3\frac{\partial A_3^+}{\partial Z}\right) = A_3^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\,d\hat{z} \\ &+ B_3^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\sin\big(2\hat{k}_3\hat{z}\big)\,d\hat{z} + A_3^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\cos\big(2\hat{k}_3\hat{z}\big)\,d\hat{z} \\ &- B_1^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z} + A_1^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z} \\ &+ B_1^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z} + A_1^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z},\end{aligned} \tag{31}$$

where the integral coefficients are defined by

$$\alpha_1 = \int_{\hat{x}=0}^{1} \hat{\omega}\epsilon_R \mathcal{L} g(n_0)\sin^2(\pi\hat{x})\,d\hat{x}, \quad \alpha_2 = \int_{\hat{x}=0}^{1} \hat{\omega}\mathcal{L}(\mathcal{V}_n+\mathcal{V}_p)n_0\sin^2(\pi\hat{x})\,d\hat{x}, \tag{32a}$$

$$\beta_1 = \int_{\hat{x}=0}^{1} \hat{\omega}\epsilon_R \mathcal{L} g(n_0)\sin(\pi\hat{x})\sin(3\pi\hat{x})\,d\hat{x}, \quad \beta_2 = \int_{\hat{x}=0}^{1} \hat{\omega}\mathcal{L}(\mathcal{V}_n+\mathcal{V}_p)n_0\sin(\pi\hat{x})\sin(3\pi\hat{x})\,d\hat{x}, \tag{32b}$$

$$\gamma_1 = \int_{\hat{x}=0}^{1} \hat{\omega}\epsilon_R \mathcal{L} g(n_0)\sin^2(2\pi\hat{x})\,d\hat{x}, \quad \gamma_2 = \int_{\hat{x}=0}^{1} \hat{\omega}\mathcal{L}(\mathcal{V}_n+\mathcal{V}_p)n_0\sin^2(2\pi\hat{x})\,d\hat{x}, \tag{32c}$$

$$\delta_1 = \int_{\hat{x}=0}^{1} \hat{\omega}\epsilon_R \mathcal{L} g(n_0)\sin^2(3\pi\hat{x})\,d\hat{x}, \quad \delta_2 = \int_{\hat{x}=0}^{1} \hat{\omega}\mathcal{L}(\mathcal{V}_n+\mathcal{V}_p)n_0\sin^2(3\pi\hat{x})\,d\hat{x}. \tag{32d}$$

The emergence of terms coupling $\sin(\pi\hat{x})$ and $\sin(3\pi\hat{x})$ profiles provides evidence of modal interaction. It has also been noted that no cross-coupling terms exist involving the $\sin(2\pi\hat{x})$ profile. This trend has been generalized to demonstrate that direct interaction is limited to any pair of $\sin(r_{odd}\pi\hat{x})$ modes, these related by their symmetry about $\hat{x}=0.5$, and any pair of $\sin(r_{even}\pi\hat{x})$ modes, the latter related by their antisymmetry about $\hat{x}=0.5$. For each lateral mode, those waves traveling along $\mathrm{O}\hat{z}$ in opposite directions also interact, this feature being described by the $\sin(2\hat{k}_r\hat{z})$ and $\cos(2\hat{k}_r\hat{z})$ integral terms above. The integrals α_1, β_1, γ_1, and δ_1 are associated with gain and absorption, while the integrals α_2, β_2, γ_2, and δ_2 describe the interaction between the electric field and charge carriers. We substitute (27) into (24) and evaluate the integral in $\hat{t}$ to obtain the second-order diffusion equation

$$\begin{aligned}\frac{\partial n_0}{\partial \tau} &= -\frac{n_0}{\mathcal{A}_n+\mathcal{A}_p} - \mathcal{B}n_0^2 - (\mathcal{C}_n+\mathcal{C}_p)n_0^3 + \frac{\mathcal{D}_n\mathcal{V}_p+\mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n+\mathcal{V}_p}\hat{\boldsymbol{\nabla}}^2 n_0 \\ &\quad - \epsilon_R\mathcal{F}g(n_0)\frac{\hat{\omega}^3}{4q_2q_3\hat{k}_1}\int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\Gamma_{ij}\sin(i\pi\hat{x})\sin(j\pi\hat{x})\right)dT,\end{aligned} \tag{33}$$

where we have

$$\begin{aligned}\Gamma_{ij} &= \cos\big((\hat{k}_i-\hat{k}_j)\hat{z}\big)\Big(A_i^+A_j^+ + B_i^+B_j^+ + A_i^-A_j^- + B_i^-B_j^-\Big) \\ &\quad + \cos\big((\hat{k}_i+\hat{k}_j)\hat{z}\big)\Big(A_i^+A_j^- + B_i^+B_j^- + A_i^-A_j^+ + B_i^-B_j^+\Big) \\ &\quad + \sin\big((\hat{k}_i-\hat{k}_j)\hat{z}\big)\Big(A_i^+B_j^+ - B_i^+A_j^+ - A_i^-B_j^- + B_i^-A_j^-\Big) \\ &\quad + \sin\big((\hat{k}_i+\hat{k}_j)\hat{z}\big)\Big(A_i^+B_j^- - B_i^+A_j^- - A_i^-B_j^+ + B_i^-A_j^+\Big).\end{aligned}$$

Before examining a partially lumped model, we present the boundary conditions and periodicity conditions that accompany (33) and the twelve secularity conditions represented by (29)–(31) and (48)–(50). Using (19) we obtain

$$\frac{\partial n_0}{\partial \hat{x}}(0,\hat{y},\hat{z},Z,\tau) = \frac{\partial n_0}{\partial \hat{x}}(1,\hat{y},\hat{z},Z,\tau) = 0. \tag{34}$$

The supply of current through the boundary of the lasing region is prescribed, where, making use of (20), we have

$$\frac{\partial n_0}{\partial \hat{y}} = \begin{cases} g_1(\hat{x},\hat{z},Z,\tau) & \text{at } \hat{y}=0, \\ g_2(\hat{x},\hat{z},Z,\tau) & \text{at } \hat{y}=1. \end{cases} \tag{35}$$

The $\hat{x}$ dependence of functions g_1 and g_2 defines a lateral current density profile which is illustrated in section 6 for two different cases. The assumption of zero normal fluxes of electrons and holes at the mirrors gives

$$\frac{\partial n_0}{\partial \hat{z}}(\hat{x},\hat{y},\hat{z},0,\tau) = \frac{\partial n_0}{\partial \hat{z}}(\hat{x},\hat{y},\hat{z},1,\tau) = 0. \tag{36}$$

As indicated in section 3, we shall approximate the imperfect mirrors in terms of the reflectivities $R^{(1)}$ at $Z = 0$ and $R^{(2)}$ at $Z = 1$. The boundary conditions for the amplitude envelopes are then given by

$$\left(A_r^+(0,T,\tau)\right)^2 = R^{(1)}\left(A_r^-(0,T,\tau)\right)^2, \quad \left(B_r^+(0,T,\tau)\right)^2 = R^{(1)}\left(B_r^-(0,T,\tau)\right)^2, \tag{37}$$

$$\left(A_r^-(1,T,\tau)\right)^2 = R^{(2)}\left(A_r^+(1,T,\tau)\right)^2, \quad \left(B_r^-(1,T,\tau)\right)^2 = R^{(2)}\left(B_r^+(1,T,\tau)\right)^2. \tag{38}$$

In addition to this the amplitude envelopes must be periodic in T such that

$$A_r^{\pm}(Z,T,\tau) = A_r^{\pm}(Z,T+2q_2q_3\hat{k}_1/\hat{\omega},\tau), \quad B_r^{\pm}(Z,T,\tau) = B_r^{\pm}(Z,T+2q_2q_3\hat{k}_1/\hat{\omega},\tau). \tag{39}$$

In conclusion, the asymptotic analysis with three lateral modes has led to a system of thirteen unknowns, comprising twelve amplitude envelopes together with the carrier density.

5. Partially lumped model.

Unlike the preceeding analysis, the partially lumped model which follows requires ad hoc averaging procedures; nevertheless, it retains lateral spatial-hole burning effects and the lateral diffusion of carriers and provides a mathematical interpretation of some of the multilateral effects that are observed in experiments.

5.1. Time-dependent equations.

We take (33) and integrate with respect to $\hat{y}$ and $\hat{z}$, giving

$$\begin{aligned}\hat{J}_y = \frac{\hat{k}_1}{2\pi q_1}\int_{\hat{y}=0}^{1}\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\Bigg(&\frac{\partial n_0}{\partial \tau} + \frac{n_0}{\mathcal{A}_n+\mathcal{A}_p} + \mathcal{B}n_0^2 + (\mathcal{C}_n+\mathcal{C}_p)n_0^3\\ &- A_x^2\,\frac{\mathcal{D}_n\mathcal{V}_p+\mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n+\mathcal{V}_p}\,\frac{\partial^2 n_0}{\partial \hat{x}^2}\\ &+ \frac{\epsilon_R\mathcal{F}\hat{\omega}^3}{4q_2q_3\hat{k}_1}\int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}}\left(g(n_0)\sum_{i=1}^{3}\sum_{j=1}^{3}\Gamma_{ij}\sin(i\pi\hat{x})\sin(j\pi\hat{x})\right)dT\Bigg)\,d\hat{z}\,d\hat{y}.\end{aligned} \tag{40}$$

The (known) current flowing through the transverse boundaries is represented by

$$\begin{aligned}\hat{J}_y &= \frac{\hat{k}_1}{2\pi q_1}\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\left[A_y^2\,\frac{\mathcal{D}_n\mathcal{V}_p+\mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n+\mathcal{V}_p}\,\frac{\partial n_0}{\partial \hat{y}}\right]_{\hat{y}=0}^{\hat{y}=1}d\hat{z}\\ &= \frac{\hat{k}_1}{2\pi q_1}\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}A_y^2\,\frac{\mathcal{D}_n\mathcal{V}_p+\mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n+\mathcal{V}_p}\left(g_2-g_1\right)d\hat{z}\end{aligned}$$

when we apply (35). We now expand n_0 in a Fourier expansion for the physically significant modes dependent on $\hat{x}$, $\hat{y}$, and $\hat{z}$ as follows:

$$\begin{aligned}n_0 = {}& N_0(\hat{x},Z,\tau) + N_1(\hat{x},Z,\tau)\cos(\hat{k}_1\hat{z}) + N_2(\hat{x},Z,\tau)\sin(\hat{k}_1\hat{z}) + N_3(\hat{x},Z,\tau)\cos(\hat{k}_2\hat{z})\\ &+ N_4(\hat{x},Z,\tau)\sin(\hat{k}_2\hat{z}) + N_5(\hat{x},Z,\tau)\cos(\hat{k}_3\hat{z}) + N_6(\hat{x},Z,\tau)\sin(\hat{k}_3\hat{z}) + \dots,\end{aligned}$$

terms in $\hat{y}$ not being listed here. We will satisfy the equations in an averaged sense, effectively making the simplifying assumption that the electron concentration is independent of $\hat{y}$ and $\hat{z}$. To achieve this aim, we truncate the Fourier series after one term. Writing $n_0 = N_0(\hat{x}, Z, \tau)$, we automatically satisfy condition (36) and obtain the electron concentration rate equation,

$$\begin{aligned}\frac{\partial N_0}{\partial \tau} &= \hat{J}_y - \frac{N_0}{\mathcal{A}_n + \mathcal{A}_p} - \mathcal{B}N_0^2 - (\mathcal{C}_n + \mathcal{C}_p)N_0^3 + A_x^2 \frac{\mathcal{D}_n\mathcal{V}_p + \mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n + \mathcal{V}_p} \frac{\partial^2 N_0}{\partial \hat{x}^2} \\ &\quad - \frac{\epsilon_R \mathcal{F}\hat{\omega}\mathcal{H}(N_0 - n^*)}{q_2 q_3 \hat{k}_1} \int_{T=0}^{2q_2q_3\hat{k}_1/\hat{\omega}} \sum_{r=1}^{3} \left(\hat{I}_r^+ + \hat{I}_r^-\right) \sin^2(r\pi\hat{x})\, dT,\end{aligned} \tag{41}$$

where we define the forward (+) and backward (−) intensities of the lasing modes to be

$$\hat{I}_r^\pm = \frac{\hat{\omega}^2}{4}\left(\left(A_r^\pm\right)^2 + \left(B_r^\pm\right)^2\right).$$

The superscript + (−) denotes those waves traveling in the direction of increasing (decreasing) $\hat{z}$. We introduce the definition of optical intensity in order to secure unique solutions: conditions (37) and (38) do not permit us to calculate a unique phase shift associated with the amplitude envelopes. Taking the secularity conditions, we obtain the photon concentration rate equations

$$\epsilon_R \hat{\omega} \frac{\partial \hat{I}_r^\pm}{\partial T} \pm \hat{k}_r \frac{\partial \hat{I}_r^\pm}{\partial Z} = 2Q_r(Z,\tau)\hat{I}_r^\pm, \tag{42}$$

in which

$$Q_r(Z,\tau) = \hat{\omega}\mathcal{L}(\epsilon_R\mathcal{H} - \mathcal{V}_n - \mathcal{V}_p) \int_{\hat{x}=0}^{1} N_0(\hat{x}, Z, \tau)\sin^2(r\pi\hat{x})\, d\hat{x} - \frac{1}{2}\hat{\omega}\mathcal{L}\epsilon_R\mathcal{H}n^*.$$

The boundary conditions are $\partial N_0/\partial\hat{x} = 0$ at $\hat{x} = 0$ and $\hat{x} = 1$ along with $\hat{I}_r^+(0, T, \tau) = R^{(1)}\hat{I}_r^-(0, T, \tau)$ and $\hat{I}_r^-(1, T, \tau) = R^{(2)}\hat{I}_r^+(1, T, \tau)$. The intensities must also satisfy $\hat{I}_r^\pm(Z, T, \tau) = \hat{I}_r^\pm(Z, T + 2q_2q_3\hat{k}_1/\hat{\omega}, \tau)$. The model has thus been simplified such that N_0 and $\hat{I}_r^\pm$ are solutions to one partial differential equation and six first-order wave equations. A dimensional form of this system, generalized for M lateral modes, is presented in Appendix B. This partially lumped model is the generalization of the traveling-wave rate equations [15] to incorporate the multilateral mode effects in Maxwell's equations. In the next subsection, we study the bifurcations associated with (41)–(42) and use these to predict the onset of multilateral mode laser operation.

5.2. Bifurcation conditions. At steady state, (42) forms a system of six first-order separable differential equations. These are solved subject to the boundary conditions $\hat{I}_r^+(0) = R^{(1)}\hat{I}_r^-(0)$ and $\hat{I}_r^-(1) = R^{(2)}\hat{I}_r^+(1)$. As a consequence, we obtain three integral conditions that correspond to each of the three lateral modes,

$$\begin{aligned}&\frac{1}{4}\hat{k}_r \ln\left(R^{(1)}R^{(2)}\right) \\ &\qquad = \frac{1}{2}\hat{\omega}\mathcal{L}\epsilon_R\mathcal{H}n^* - \hat{\omega}\mathcal{L}(\epsilon_R\mathcal{H} - \mathcal{V}_n - \mathcal{V}_p) \int_0^1 \int_0^1 N_0(\hat{x}, Z)\sin^2(r\pi\hat{x})\, d\hat{x}\, dZ.\end{aligned} \tag{43}$$

Substituting the solutions for the forward and backward intensities into the steady-state problem from (41) leaves us with the following nonlinear ordinary differential equation to be solved numerically:

$$\begin{aligned} A_x^2\,\frac{\mathcal{D}_n\mathcal{V}_p+\mathcal{D}_p\mathcal{V}_n}{\mathcal{V}_n+\mathcal{V}_p}\frac{\partial^2 N_0}{\partial\hat{x}^2}-\frac{N_0}{\mathcal{A}_n+\mathcal{A}_p}-\mathcal{B}N_0^2-(\mathcal{C}_n+\mathcal{C}_p)N_0^3+\hat{J}_y \\ =\frac{1}{2}\,\epsilon_R\hat{\omega}^2\mathcal{F}\mathcal{H}(N_0-n^*)\sum_{r=1}^{3}\sin^2(r\pi\hat{x})f_r(Z), \end{aligned}\tag{44}$$

in which the functions $f_r(Z)=4\big(\hat{I}_r^+(Z)+\hat{I}_r^-(Z)\big)/\hat{\omega}^2$ are written in the form

$$\begin{aligned} C_r\Bigg(\exp\left\{\frac{2\hat{\omega}\mathcal{L}(\epsilon_R\mathcal{H}-\mathcal{V}_n-\mathcal{V}_p)}{\hat{k}_r}\int_{Z'=0}^{Z}G(Z')\,dZ'-\frac{\hat{\omega}\mathcal{L}\epsilon_R\mathcal{H}n^*Z}{\hat{k}_r}\right\} \\ +\sqrt{\frac{R^{(2)}}{R^{(1)}}}\exp\left\{\frac{2\hat{\omega}\mathcal{L}(\epsilon_R\mathcal{H}-\mathcal{V}_n-\mathcal{V}_p)}{\hat{k}_r}\int_{Z'=Z}^{1}G(Z')\,dZ'+\frac{\hat{\omega}\mathcal{L}\epsilon_R\mathcal{H}n^*(Z-1)}{\hat{k}_r}\right\}\Bigg), \end{aligned}$$

where

$$G(Z')=\int_{\hat{x}=0}^{1}N_0(\hat{x},Z')\sin^2(r\pi\hat{x})\,d\hat{x}.$$

From (34) we have the no flux boundary conditions

$$\frac{\partial N_0}{\partial\hat{x}}=0\qquad\text{at}\quad\hat{x}=0,1.\tag{45}$$

The additional unknowns, C_r, are constants of integration that emerge when calculating the intensities and are determinable through solving the electrical-optical problem (43)–(45) with prescribed current density $\hat{J}_y$.

Experiments suggest that the fundamental mode reaches threshold first, since it couples most effectively to the injected carrier distribution. The current is then increased until the gain distribution begins to couple effectively to the next mode. At such a threshold current the next mode begins to share power with the fundamental mode [6]. A series of bifurcation conditions arising from (43)–(45) describe this phenomenon, allowing us to calculate the threshold current density at which each mode begins lasing (note that the corresponding current is given by the integral of the current density, $\hat{J}_y$, over the area $0\le Z\le 1$, $0\le\hat{x}\le 1$):

(i) *First threshold.* For both forms of $\hat{J}_y$ considered below, we ascertain numerically that the $r=1$ lateral mode is activated first. At a current equal to the first threshold there is zero leading-order intensity, that is, $C_1=C_2=C_3=0$. The unknown threshold current density $\hat{J}_y$ and threshold electron concentration N_0 satisfy (43) with $r=1$ and (44)–(45). This constitutes the first bifurcation condition and determines the threshold current at which the dominant ($\sin(\pi\hat{x})$) mode is switched on.

(ii) *Next threshold.* As the $\sin(\pi\hat{x})$ mode is already activated, we take $C_2=C_3=0$ and include the unknown C_1 in (44). Either the $\sin(2\pi\hat{x})$ mode or the $\sin(3\pi\hat{x})$ mode may be switched on next, depending on which of the criteria below results in the lowest threshold current:

(a) $\sin(2\pi\hat{x})$. The unknown threshold current density $\hat{J}_y$, the threshold electron concentration N_0, and C_1 satisfy (43) with $r=1$ and $r=2$ and (44)–(45).

(b) $\sin(3\pi\hat{x})$. The unknown threshold current density $\hat{J}_y$, the threshold electron concentration N_0, and C_1 satisfy (43) with $r = 1$ and $r = 3$ and (44)–(45).

System (a) or (b) constitutes the second bifurcation condition. We thus have a physical criterion for the onset of multilateral mode operation, a dimensional form of which is presented in Appendix C. A final bifurcation condition, of the same ilk as the two described above, determines the threshold current density at which the remaining lateral mode begins lasing.

One consequence of this multimode formulation is a changing near-field lateral intensity envelope. Experiments demonstrate that this evolution varies depending on the form of the semiconductor laser and in particular on the lateral width of the active region. For narrow lateral widths, the single peak that exists at the center of the interval becomes far more pronounced with increased current [1, p. 66]. The 10–20μm stripe lasers measured in [13, p. 377] have an associated near-field distribution which broadens and then splits into two individual peaks as a new mode is activated. This particular observation compares closely to the results obtained in the following section. In broad-area semiconductor lasers a rippled profile has been observed, this being interpreted as the superposition of several lateral mode profiles [9]. As the pumped current increases, the near-field patterns exhibit a greater number of lobes, suggesting that more and more lateral modes are being activated. This model, albeit dealing with a relatively simple case, predicts this effect, enabling us to calculate the current density at which each lateral mode becomes operational.

6. Numerical results.

6.1. Introduction. Solutions to the steady-state electrical-optical problem (43)–(45) are split into two cases. The current density varies on the lateral length-scale and is expressed in the form $\hat{J}_y = J_{AMP}F(\hat{x})$ for $0 \leq Z \leq 1$, where J_{AMP} and $F(\hat{x})$ specify the current density amplitude and profile, respectively. We consider applications A and B which deal with two different current density distributions (see Figure 2), labeled $F_A(\hat{x})$ and $F_B(\hat{x})$, respectively. The latter case deals with a comparatively broader distribution. Varying $F(\hat{x})$ is found to impact the order in which the three lateral modes are activated. The closed problem (43)–(45) is solved using a finite-difference scheme. All of the results presented below are based on the parameter values given in Table 1.

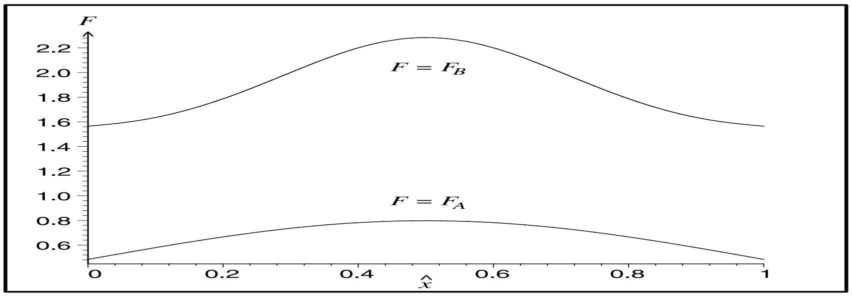


FIG. 2. *Graph displaying the two current density distributions used with $F(\hat{x}) = F_A(\hat{x})$ and $F(\hat{x}) = F_B(\hat{x})$ explored in cases A and B, respectively.*

6.2. Case study A: $\boldsymbol{F = F_A(\hat{x})}$. In this first case, the $\sin(r\pi\hat{x})$ lateral modes are activated in the order $r = 1, 2, 3$ as the pumped current increases. The corresponding threshold values of J_{AMP} are $t_1 \sim 26$, $t_2 \sim 34$, and $t_3 \sim 36$, respectively.

6.2.1. Stable and unstable steady states. In this section, a measure of the power, P_r, attributed to each lateral mode is found by averaging the backward intensity over the longitudinal length,

$$P_r = \int_{Z=0}^{1} \hat{I}_r^-(Z)\,dZ. \tag{46}$$

Time-independent numerical simulations determine a number of steady-state solutions, not all of which are stable. To demonstrate this, we analyze the power-current graphs belonging to the $r = 1$ and $r = 2$ lateral modes. In particular, there are transcritical bifurcations at those J_{AMP} values at which each lateral mode is switched on (see Figure 3). Prior to the first current threshold, the stable steady state has zero power (see branches A and D for $0 \leq J_{AMP} < t_1$ in Figure 3). At the first threshold, a continuation of zero power remains a valid solution of the equations. However, this solution is unstable following the activation of the $\sin(\pi\hat{x})$ mode. Thus, for $t_1 < J_{AMP} < t_2$, branches B and D represent the stable solution. An identical situation arises when the second threshold current is encountered. Here the $\sin(2\pi\hat{x})$ mode becomes operational. Beyond this point it is possible for the solution to continue along branches B and D; however, at the second threshold this solution too becomes unstable. A new stable solution exists due to the presence of the second lateral mode (see branches C and E for $J_{AMP} > t_2$ in Figure 3). The drop in gradient about $J_{AMP} = t_2$ in the top graph in Figure 3 symbolizes a reduction in the efficiency of the $\sin(\pi\hat{x})$ mode due to the onset of competition for photons (see subsection 6.2.2). These trends continue as the third, $\sin(3\pi\hat{x})$, mode is activated.

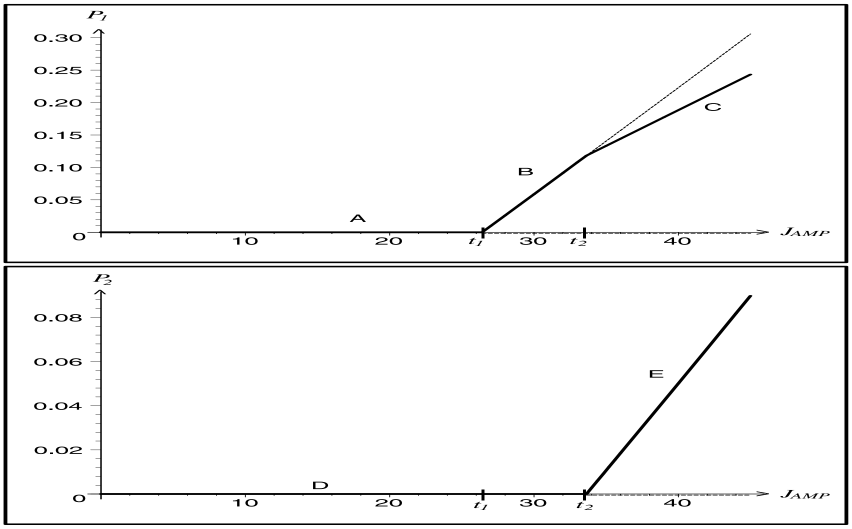


FIG. 3. *Power-current graphs for the $r = 1$ (top) and $r = 2$ (bottom) lateral modes. Solid and dashed lines indicate the stable and unstable steady states, respectively. The displayed quantities are dimensionless.*

6.2.2. Competition for photons. Hence, using (46) we are able to track the power associated with each of the three modes as the current density amplitude rises beyond each calculated threshold (see Figure 4). At the second and third threshold currents (corresponding to the broken vertical lines) there is a drop in gradient asso-

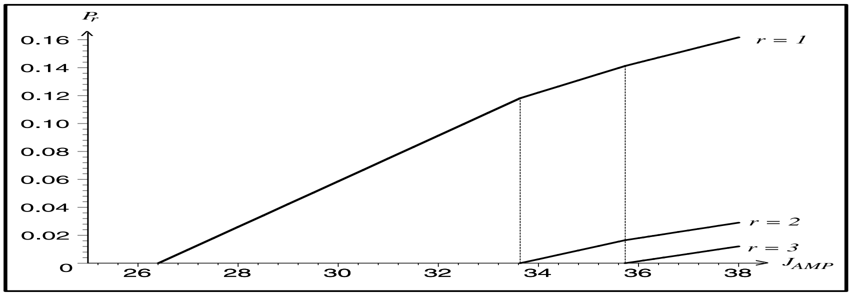


FIG. 4. *Graph showing the relationship between power and current with* $r = 1, 2, 3$ *representing the* $\sin(\pi\hat{x})$, $\sin(2\pi\hat{x})$, *and* $\sin(3\pi\hat{x})$ *modes, respectively. The values plotted are dimensionless.*

ciated with those modes already operating. This reduction in efficiency is caused by the competition for photons and the subsequent sharing of the available gain.

6.2.3. Lateral spatial-hole burning and changing near-field patterns. The lateral near-field intensity profile evolves following the activation of each mode, due to the superposition of the three $\sin(r\pi\hat{x})$ curves. The former is defined by

$$L^2(\hat{x}) = \left(\sum_{r=1}^{3} \sqrt{\left(\hat{I}_r^+(0) + \hat{I}_r^-(0)\right)}\,\sin(r\pi\hat{x})\right)^2 \tag{47}$$

and has been plotted in Figure 5 for case A. As the current increases, there is a shift in the peak of the optical distribution to one side, an effect observed in some of the stripe lasers studied in [13, p. 377]. This case does not generate a multilobed lateral near-field profile. Instead, the onset of the third ($\sin(3\pi\hat{x})$) mode creates a single ripple (top curve in Figure 5), this feature also appearing within the near-field scans in [1, p. 66]. The lateral intensity of the $\sin(\pi\hat{x})$ mode produces the high local rate of stimulated recombination at the center of the active region (bottom curve in Figure 5). After a sufficient increase in current, this causes the curvature at the peak of the carrier distribution to become inverted, forming a dip in the profile [13, p. 380]. This is seen in Figure 6, where we have plotted the lateral variation of N_0 at $Z = 0.5$. Note that the evolution of the lateral carrier profile at higher currents reflects a pattern of optical modes [8, p. 585], which is further illustrated in case B. The phenomenon described here is known as spatial-hole burning and takes place in the lateral direction in this particular problem. Note that longitudinal spatial-hole burning properties have been neglected due to the omission from N_0 of $\hat{z}$ dependence.

6.3. Case study B: $\boldsymbol{F = F_B(\hat{x})}$. In this second case, the sequence of activation of the $\sin(r\pi\hat{x})$ modes is given by $r = 1, 3, 2$; the respective threshold current density amplitudes are given by $J_{AMP} \sim 9.5$, 12.6, and 13.1. It follows that the physical behavior differs as a result of this revised sequence (or lateral mode hop). In particular, the near-field profile ($L^2(\hat{x})$) develops a lobed effect that is similar to that presented in [13, p. 377]. The distribution of light gradually focuses around two particular points along the active layer's lateral length (see Figure 7). The spatial-hole burning mechanism affects the lateral carrier profile quite dramatically, as indicated in Figure 8, where N_0 is plotted for $Z = 0.5$. At higher currents, a series of peaks and troughs like that in Figure 8 is typical of broad-area semiconductor lasers (see,

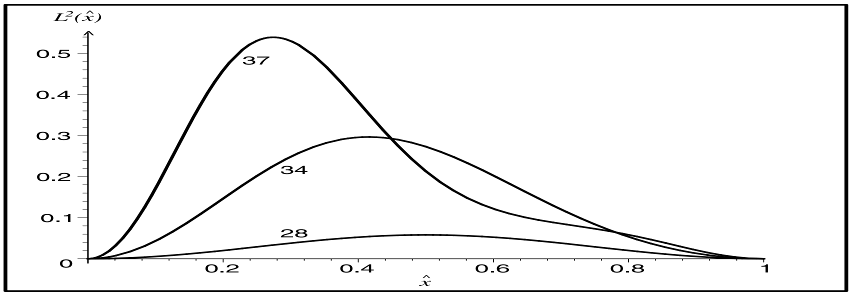


FIG. 5. *Tracing the near-field lateral intensity profile with increasing current in case A. The labels correspond to the values of J_{AMP} taken in each computation. All values used are nondimensional.*

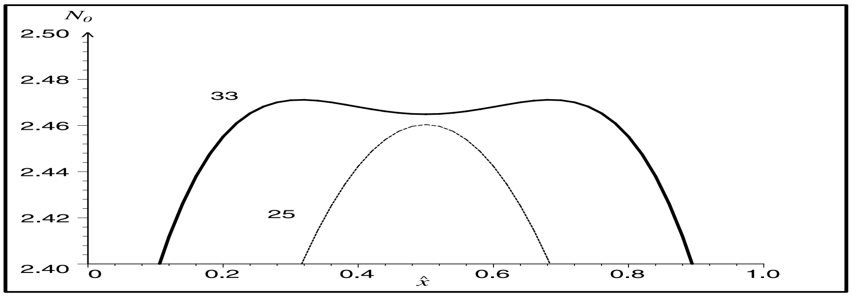


FIG. 6. *Zooming in on the center of the lateral carrier profile at $Z = 0.5$ to highlight the onset of lateral spatial-hole burning. The labels on each curve refer to the increasing dimensionless current density amplitude J_{AMP}.*

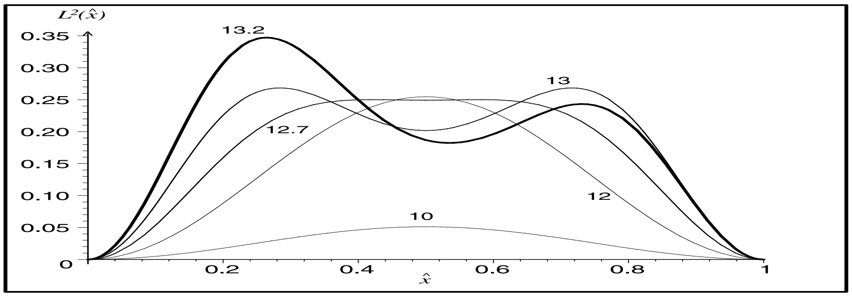


FIG. 7. *Tracing the near-field lateral intensity profile with increasing current in case B. Again the labels correspond to the increasing J_{AMP}. All values used are nondimensional.*

for example, [13, pp. 380–381]). Note that the increased levels of stimulated gain, which occur at the peaks in Figure 7, coincide with a local reduction in the electron concentration. As the $\sin(3\pi\hat{x})$ lateral mode is activated, the electron concentration can be seen to exhibit a $\cos(6\pi\hat{x})$ mode.

7. Conclusions. The models discussed in this paper have been employed to simulate electrical-optical effects in the active region of a multilateral mode semiconductor laser. The starting point was a system based on Maxwell's equations together with models for the polarization, charge density, and current density (which represent a realistic semiconductor medium). A multiple-scale asymptotic expansion resulted

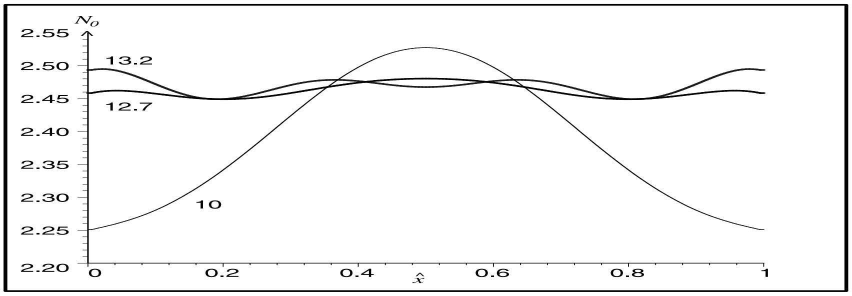


FIG. 8. *Lateral variation of the electron density at* $Z = 0.5$ *when one (*$J_{AMP} \sim 10$*), two (*$J_{AMP} \sim 12.7$*), and then three (*$J_{AMP} \sim 13.2$*) modes are operating in case B.*

in a leading-order problem in which the linear equations for the lasing mode can be viewed as representing a waveguide. The infinite number of possible lateral mode shapes in the waveguide has been truncated to the three most commonly observed in experiments. At the next order, a set of twelve nonlinear secularity conditions model the lasing mode envelopes. The original system has been reduced to twelve first-order wave equations for the envelope of the lasing mode and one second-order diffusion equation for the electron concentration.

Averaging over the transverse and short longitudinal length-scales led to the partially lumped model, which consists of six first-order wave equations and one nonlinear (lateral) diffusion equation for the electron concentration. The traveling-wave rate equations form a counterpart to the single lateral mode version of this partially lumped model. However, in the existing literature, there are no counterparts to the multilateral partially lumped model or criterion for the onset of multimode operation, these representing the most important aspects of the current paper.

We conclude by noting that the simplified models provide a basis for the study of broad-area lasers. Ongoing research incorporates a pair of two-dimensional heat equations into the partially lumped model in order to simulate thermal hot-spots in high-power lasers (see, for example, [4] and [5]).

Appendix A. Secularity conditions for amplitude envelopes $B_r^+(Z,T,\tau)$. The first-order wave equations associated with B_1^+, B_2^+, and B_3^+ are presented with the integral terms α_1, α_2, β_1, β_2, γ_1, γ_2, δ_1, and δ_2 as defined in (32a)–(32d):

(48)

$$\frac{2\pi q_1}{\hat{k}_1}\left(\epsilon_R\hat{\omega}\frac{\partial B_1^+}{\partial T} + \hat{k}_1\frac{\partial B_1^+}{\partial Z}\right) = B_1^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1 - \alpha_2\big)\,d\hat{z}$$

$$- A_1^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1-\alpha_2\big)\sin\big(2\hat{k}_1\hat{z}\big)\,d\hat{z} + B_1^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\alpha_1-\alpha_2\big)\cos\big(2\hat{k}_1\hat{z}\big)\,d\hat{z}$$

$$- A_3^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z} + B_3^+\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z}$$

$$- A_3^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z} + B_3^-\int_{\hat{z}=0}^{2\pi q_1/\hat{k}_1}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z},$$

$$\begin{aligned}(49)\quad &\frac{2\pi q_2}{\hat{k}_2}\left(\epsilon_R\hat{\omega}\frac{\partial B_2^+}{\partial T}+\hat{k}_2\frac{\partial B_2^+}{\partial Z}\right)=B_2^+\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\,d\hat{z}\\ &-A_2^-\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\sin\big(2\hat{k}_2\hat{z}\big)\,d\hat{z}+B_2^-\int_{\hat{z}=0}^{2\pi q_2/\hat{k}_2}\big(\gamma_1-\gamma_2\big)\cos\big(2\hat{k}_2\hat{z}\big)\,d\hat{z},\end{aligned}$$

(50)

$$\begin{aligned}&\frac{2\pi q_3}{\hat{k}_3}\left(\epsilon_R\hat{\omega}\frac{\partial B_3^+}{\partial T}+\hat{k}_3\frac{\partial B_3^+}{\partial Z}\right)=B_3^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\,d\hat{z}\\ &-A_3^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\sin\big(2\hat{k}_3\hat{z}\big)\,d\hat{z}+B_3^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\delta_1-\delta_2\big)\cos\big(2\hat{k}_3\hat{z}\big)\,d\hat{z}\\ &+A_1^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z}+B_1^+\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1-\hat{k}_3)\hat{z}\big)\,d\hat{z}\\ &-A_1^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\sin\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z}+B_1^-\int_{\hat{z}=0}^{2\pi q_3/\hat{k}_3}\big(\beta_1-\beta_2\big)\cos\big((\hat{k}_1+\hat{k}_3)\hat{z}\big)\,d\hat{z}.\end{aligned}$$

Appendix B. Dimensional traveling-wave rate equations for multilateral mode lasers. This appendix presents the traveling-wave rate equations for a broad-area semiconductor laser with M lateral modes. Letting k_r represent the dimensional wave number associated with the $\sin(r\pi x/w)$ lateral mode, the dimensional dispersion relation is given by $k_r^2=\omega^2\epsilon\mu_0-(r\pi/w)^2$. The current density profile is denoted by $J(x)$ below. After transforming to dimensional variables in (41)–(42), we arrive at the following diffusion equation for the electron concentration $n(x,z,t)$:

$$\begin{aligned}(51)\quad &\frac{\partial n}{\partial t}=\frac{J(x)}{ed}-\frac{n}{A_n+A_p}-Bn^2-(C_n+C_p)n^3+\left(\frac{D_n\mu_p+D_p\mu_n}{\mu_n+\mu_p}\right)\frac{\partial^2 n}{\partial x^2}\\ &-2a(n-n_t)\sum_{r=1}^{M}\left(I_r^++I_r^-\right)\sin^2\!\left(\frac{r\pi x}{w}\right).\end{aligned}$$

We deduce that the optical intensities, $I_r^\pm(z,t)$, satisfy

$$(52)\quad \epsilon\omega\frac{\partial I_r^\pm}{\partial t}\pm\frac{k_r}{\mu_0}\frac{\partial I_r^\pm}{\partial z}=\left(\frac{2\omega e}{w}\left(\frac{\epsilon a}{e}-\mu_n-\mu_p\right)\int_0^w n\sin^2\!\left(\frac{r\pi x}{w}\right)dx-\omega\epsilon a n_t\right)I_r^\pm,$$

where $r=1,2,\dots,M$. The boundary conditions are $I_r^+(0,t)=R^{(1)}I_r^-(0,t)$ and $I_r^-(L,t)=R^{(2)}I_r^+(L,t)$ along with $\partial n/\partial x=0$ at $x=0$ and $x=w$.

Appendix C. A dimensional criterion for multimode operation. This appendix presents a dimensional criterion that corresponds to the onset of multimode laser operation. The fundamental mode, characterized by the lateral profile $\sin(\pi x/w)$, is activated at the first threshold current. At the second threshold current density amplitude, J_P, the $\sin(p\pi x/w)$ lateral mode is switched on and multimode lasing is established. Note that the particular mode to be activated next (where $p\geq 2$) depends upon the lateral current density profile, $\bar{F}(x)$, and the bifurcation condition, as seen in section 6. The bifurcation condition thus consists of (53)–(54) and (55) for $r=1$ and $r=p$, where we solve for the threshold electron concentration $n(x,z)$, the constant of integration $\bar{C}_1$ (this corresponding to the dominant mode), and J_P. The dimensional ordinary differential equation for the threshold carrier density takes the form

$$\frac{D_n\mu_p + D_p\mu_n}{\mu_n+\mu_p}\frac{\partial^2 n}{\partial x^2} + J_P\bar{F}(x) = \frac{n}{A_n+A_p} + Bn^2 + (C_n+C_p)n^3 + 2a(n-n_t)\sin^2\Big(\frac{\pi x}{w}\Big)\bar{f}_1(z), \tag{53}$$

where the sum of the forward and backward intensities of the $\sin(\pi x/w)$ lateral mode is represented by

$$\bar{f}_1(z) = \bar{C}_1\left(\exp\left\{\frac{2\omega\mu_0 e}{wk_1}\Big(\frac{\epsilon a}{e}-\mu_n-\mu_p\Big)\int_{z'=0}^{z}\bar{G}(z')\,dz' - \frac{\omega\mu_0\epsilon a n_t}{k_1}\,z\right\} + \sqrt{\frac{R^{(2)}}{R^{(1)}}}\exp\left\{\frac{2\omega\mu_0 e}{wk_1}\Big(\frac{\epsilon a}{e}-\mu_n-\mu_p\Big)\int_{z'=z}^{L}\bar{G}(z')\,dz' - \frac{\omega\mu_0\epsilon a n_t}{k_1}\,(L-z)\right\}\right),$$

in which

$$\bar{G}(z') = \int_{x=0}^{w} n(x,z')\sin^2\Big(\frac{\pi x}{w}\Big)\,dx.$$

We apply the no flux condition

$$\frac{\partial n}{\partial x} = 0 \quad \text{at } x = 0, w. \tag{54}$$

The criterion is completed by the following integral condition:

$$\frac{1}{4}k_r\ln\Big(R^{(1)}R^{(2)}\Big) = \frac{1}{2}\omega\mu_0\epsilon a n_t L - \frac{\omega\mu_0 e}{w}\Big(\frac{\epsilon a}{e}-\mu_n-\mu_p\Big)\int_0^L\int_0^w n\sin^2\Big(\frac{r\pi x}{w}\Big)\,dx\,dz. \tag{55}$$

REFERENCES

[1] G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*, Van Nostrand Reinhold, New York, 1993.

[2] S. Beri, M. Yousefi, P. C. De Jager, D. Lenstra, and M. K. Smit, *Complete rate equation modelling for the dynamics of multi-mode semiconductor lasers*, in Proceedings of the IEEE/LEOS Benelux Chapter Symposium, Eindhoven, The Netherlands, 2006, pp. 141–144.

[3] W. W. Chow, S. W. Koch, and M. Sargent, *Semiconductor-Laser Physics*, Springer-Verlag, Berlin, 1994.

[4] B. P. Cox, *New Models for Multilateral Mode Semiconductor Lasers*, Ph.D. thesis, The University of Birmingham, Birmingham, UK, 2006.

[5] B. P. Cox and W. R. Smith, *Predictions of thermoelastic stress in a broad-area semiconductor laser*, Appl. Phys. Lett., 90 (2007), 121105.

[6] B. W. Hakki, *GaAs double heterostructure lasing behaviour along the junction plane*, J. Appl. Phys., 46 (1975), pp. 292–302.

[7] H. F. Hofmann and O. Hess, *Quantum Maxwell-Bloch equations for spatially inhomogeneous semiconductor lasers*, Phys. Rev. A, 59 (1999), pp. 2342–2358.

[8] H. Kressel and J. K. Butler, *Semiconductor Lasers and Heterojunction LEDs*, Academic Press, New York, 1977.

[9] R. J. Lang, A. G. Larsson, and J. G. Cody, *Lateral modes of broad area semiconductor lasers: Theory and experiment*, IEEE J. Quantum Electron., 27 (1991), pp. 312–320.

[10] A. C. Newell and J. V. Moloney, *Nonlinear Optics*, Addison–Wesley, Redwood City, CA, 1991.

[11] K. Petermann, *Laser Diode Modulation and Noise*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[12] W. R. Smith, J. R. King, and B. Tuck, *Mathematical modelling of electrical-optical effects in semiconductor laser operation*, SIAM J. Appl. Math., 61 (2001), pp. 2122–2147.

[13] G. H. B. Thompson, *Physics of Semiconductor Laser Devices*, John Wiley and Sons, Chichester, UK, 1980.

[14] R. K. Wangsness, *Electromagnetic Fields*, John Wiley and Sons, New York, 1979.

[15] J. Z. Wilcox and L. W. Casperson, *Power characteristics of single-mode semiconductor lasers*, J. Appl. Phys., 56 (1984), pp. 57–64.

[16] D. P. Wilt and A. Yariv, *A self-consistent static model of the double-heterostructure laser*, IEEE J. Quantum Electron., 17 (1981), pp. 1941–1949.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 22–36

MULTIFREQUENCY TRANS-ADMITTANCE SCANNER: MATHEMATICAL FRAMEWORK AND FEASIBILITY[*]

SUNGWHAN KIM[†], JEEHYUN LEE[‡], JIN KEUN SEO[‡], EUNG JE WOO[§], AND HABIB ZRIBI[§]

Abstract. A trans-admittance scanner (TAS) is a device for breast cancer diagnosis based on numerous experimental findings that complex conductivities of breast tumors significantly differ from those of surrounding normal tissues. In TAS, we apply a sinusoidal voltage between a hand-held electrode and a scanning probe placed on the breast skin to make current travel through the breast. The scanning probe has an array of electrodes at zero voltage. We measure exit currents (Neumann data) through the electrodes that provide a map of trans-admittance data over the breast surface. The inverse problem of TAS is to detect a suspicious abnormality underneath the breast skin from the measured Neumann data. Previous anomaly detection methods used the difference between the measured Neumann data and a reference Neumann data obtained beforehand in the absence of anomaly. However, in practice, the reference data is not available and its computation is not possible since the inhomogeneous complex conductivity of the normal breast is unknown. To deal with this problem, we propose a frequency-difference TAS (fdTAS), in which a weighted frequency difference of the trans-admittance data measured at a certain moment is used for anomaly detection. This paper provides a mathematical framework and the feasibility of fdTAS by showing the relationship between the anomaly information and the weighted frequency difference of the Neumann data.

Key words. breast cancer detection, electrical conductivity, T-Scan, anomaly estimation algorithm

AMS subject classifications. 35R30, 34A45, 65N21, 78A30, 78A70

DOI. 10.1137/070683593

1. Introduction.

A trans-admittance scanner (TAS) is a device for breast cancer diagnosis that is based on the consensus that complex conductivity values of breast tumors significantly differ from those of surrounding normal tissues [4, 18, 22, 35, 37]. For example, T-Scan is a commercially available TAS system that has been suggested for adjunctive clinical uses with X-ray mammography to decrease equivocal findings and thereby reduce unnecessary biopsies [4]. In TAS, a patient holds a reference electrode with one hand through which a sinusoidal voltage $V_0 \sin \omega t$ is applied, while a scanning probe at the ground potential is placed on the surface of the breast. The voltage difference $V_0 \sin \omega t$ produces electric current flowing through the breast region. See Figure 1. The resulting electric potential at a position $x = (x_1, x_2, x_3)$ and time t can be expressed as the real part of $u(x)e^{i\omega t}$, where the complex potential $u(x)$ is governed by the equation $\nabla \cdot ((\sigma + i\omega\epsilon)\nabla u(x)) = 0$ in the subject, where σ and ϵ denote the conductivity and permittivity, respectively. The scanning probe is equipped with a planar array of electrodes, and we measure exit currents (Neumann

[*] Received by the editors February 24, 2007; accepted for publication (in revised form) May 2, 2008; published electronically July 25, 2008. This work was supported by the SRC/ERC program of MOST/KOSEF (R11-2002-103) and the Basic Research Program of KOSEF (R01-2005-000-10339-0).
http://www.siam.org/journals/siap/69-1/68359.html

[†] Division of Liberal Arts, Hanbat National University, Daejeon, 305-719, Republic of Korea (sungwhan@hanbat.ac.kr).

[‡] Department of Mathematics, Yonsei University, Seoul 120-749, Republic of Korea (ezhyun@yonsei.ac.kr, seoj@yonsei.ac.kr).

[§] College of Electronics and Information, Kyung Hee University, Suwon 449-701, Republic of Korea (ejwoo@khu.ac.kr, zribi@cmapx.polytechnique.fr).

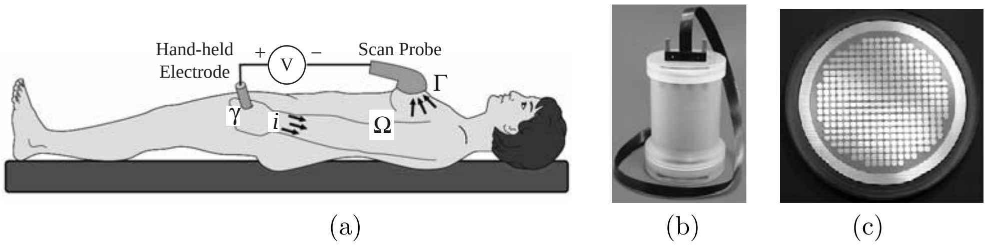


FIG. 1. *TAS setup.* (a) *Voltage is applied between the hand-held electrode and the planar array of electrodes in the scan probe. Exit currents through the scan probe are measured to provide trans-admittance data.* (b) *Picture of a scan probe and* (c) *its electrode array.*

data) $g = -(\sigma + i\omega\epsilon)\frac{\partial u}{\partial \mathbf{n}}$ which reflect electrical properties of tissues under the scan probe. Here, $\frac{\partial u}{\partial \mathbf{n}}$ is the normal derivative of u.

The inverse problem of TAS is to detect a suspicious abnormality in a breast region underneath the probe from measured Neumann data g. All previous anomaly detection methods utilize a difference $g - g^*$, where g^* is a reference Neumann data measured beforehand without any anomaly inside the breast region [1, 34]. This difference $g - g^*$ can be viewed as a kind of background subtraction so that it makes the anomaly apparently visible. However, in practice, it is not available in most cases, and calculating g^* is not possible since the inhomogeneous complex conductivity of a specific normal breast is unknown. In order for TAS to be more practical, we should avoid using this background difference data $g - g^*$. Therefore, we propose a frequency difference TAS method which uses a frequency difference of trans-admittance data measured at a certain moment.

To be precise, let the human body occupy a three-dimensional domain Ω with a smooth boundary $\partial\Omega$. Let Γ and γ be portions of $\partial\Omega$, denoting the probe plane placed on the breast and the surface of the metallic reference electrode, respectively. Through γ, we apply a sinusoidal voltage of $V_0 \sin \omega t$ with its frequency $f = \omega/2\pi$ in a range of 50 Hz to 500 kHz. Then the corresponding complex potential u_ω at ω satisfies the following mixed boundary value problem:

$$(1) \qquad \begin{cases} \nabla \cdot ((\sigma + i\omega\epsilon)\nabla u_\omega(x)) = 0 \quad \text{in } \Omega, \\ u_\omega(x) = 0, \quad x \in \Gamma, \\ u_\omega(x) = V_0, \quad x \in \gamma, \\ (\sigma + i\omega\epsilon)\nabla u_\omega(x) \cdot \mathbf{n}(x) = 0, \quad x \in \partial\Omega \setminus (\Gamma \cup \gamma), \end{cases}$$

where $\mathbf{n}$ is the unit outward normal vector to the boundary $\partial\Omega$. Note that both $\sigma = \sigma(x, \omega)$ and $\epsilon = \epsilon(x, \omega)$ depend on ω. The scan probe Γ consists of a planar array of electrodes $\mathcal{E}_1, \ldots, \mathcal{E}_m$, and we measure exit current $g_\omega(j)$ through each electrode $\mathcal{E}_j$:

$$g_\omega(j) := -\int_{\mathcal{E}_j} (\sigma + i\omega\epsilon)\, \nabla u_\omega \cdot \mathbf{n}\, ds \qquad (j = 1, \ldots, m),$$

where ds is the area element.

In the frequency-difference TAS (fdTAS), we apply voltage with two different frequencies $f_1 = \omega_1/2\pi$ and $f_2 = \omega_2/2\pi$ with 50 Hz $\leq f_1 < f_2 \leq$ 500 kHz and measure two sets of corresponding Neumann data g_{ω_1} and g_{ω_2} through Γ at the same time. We assume that there exists a region of breast tumor D beneath the probe Γ so that $\sigma + i\omega\epsilon$ changes abruptly across ∂D. (See Remark 2.1.) The inverse problem of fdTAS is to detect the anomaly D beneath Γ from a difference between g_{ω_1} and g_{ω_2}.

In order for any detection algorithm to be practical, we must take into account the following limitations:

(a) Since Ω differs for each subject, the algorithm should be robust against any change in the geometry of Ω and also any change in the complex conductivity distribution outside the breast region.

(b) The Neumann data g_ω is available only on a small surface Γ instead of the whole surface $\partial\Omega$.

(c) Since the inhomogeneous complex conductivity of the normal breast without D is unknown, it is difficult to obtain the reference Neumann data g^*_ω in the absence of D.

These limitations are indispensable to a TAS model in practical situations, and these are the reasons why we try to improve the previous techniques [1, 2, 3, 6, 7, 8, 9, 10, 11, 13, 17, 21, 24, 25, 26, 27, 28, 31, 34] by using frequency difference.

In the fdTAS model, we use a weighted frequency difference of Neumann data $g_{\omega_2} - \alpha g_{\omega_1}$ instead of $g_{\omega_2} - g_{\omega_1}$. The weight constant α is approximately $\alpha \approx \frac{\int_\Gamma g_{\omega_2} ds}{\int_\Gamma g_{\omega_1} ds}$, and the weight is a crucial factor in the anomaly detection. We should note that the simple difference $g_{\omega_2} - g_{\omega_1}$ may fail to extract the anomaly due to the complicated structure of the solution of the complex conductivity equation. See Remark 3.3. In Theorem 3.2, we explain how $g_{\omega_2} - \alpha g_{\omega_1}$ reflects a contrast in complex conductivity values between the anomaly D and surrounding normal tissues. The approximate representation formula is given in Remark 3.4.

Recently, we published a preliminary experimental validation study of fdTAS in [32]. However, this previous work lacks a mathematical analysis of the method. We therefore describe a rigorous mathematical framework of the fdTAS method in this paper.

2. Mathematical model and the feasibility of fdTAS.

We assume that σ and ϵ are isotropic, positive, and piecewise smooth functions in $\overline{\Omega}$. Let u_ω be the $H^1(\Omega)$-solution of (1). Denoting the real and imaginary parts of u_ω by $v_\omega = \Re u_\omega$ and $h_\omega = \Im u_\omega$, the mixed boundary value problem (1) can be expressed as the following coupled system:

$$(2)\qquad \begin{cases} \nabla\cdot(\sigma\nabla v_\omega) - \nabla\cdot(\omega\epsilon\nabla h_\omega) = 0 & \text{in} \quad \Omega, \\ \nabla\cdot(\omega\epsilon\nabla v_\omega) + \nabla\cdot(\sigma\nabla h_\omega) = 0 & \text{in} \quad \Omega, \\ \quad v_\omega = 0 \quad \text{and} \quad h_\omega = 0 & \text{on} \quad \Gamma, \\ \quad v_\omega = V_0 \quad \text{and} \quad h_\omega = 0 & \text{on} \quad \gamma, \\ \mathbf{n}\cdot\nabla v_\omega = 0 \quad \text{and} \quad \mathbf{n}\cdot\nabla h_\omega = 0 & \text{on} \partial\Omega \setminus (\Gamma\cup\gamma). \end{cases}$$

The measured Neumann data g_ω can be decomposed into

$$g_\omega(x) := \underbrace{\mathbf{n}\cdot(-\sigma\nabla v_\omega(x) + \omega\epsilon\nabla h_\omega(x))}_{\text{real part}} + i\ \underbrace{\mathbf{n}\cdot(-\sigma\nabla h_\omega(x) - \omega\epsilon\nabla v_\omega(x))}_{\text{imaginary part}}, \qquad x\in\Gamma.$$

The solution of the coupled system (2) is a kind of saddle point [5, 12], and we have the following relations:

$$(3)\qquad V_0\int_\Gamma \Re(g_\omega)ds = \min_{v\in\mathcal{H}_{re}}\max_{h\in\mathcal{H}_{im}}\int_\Omega \left[\sigma|\nabla v|^2 - 2\omega\epsilon\nabla v\cdot\nabla h - \sigma|\nabla h|^2\right]dx$$

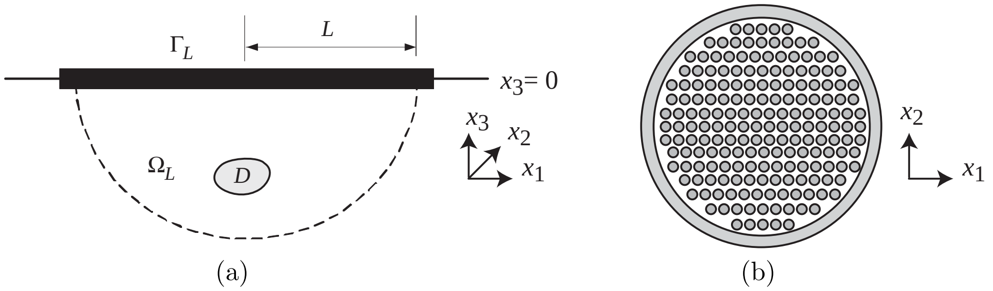


FIG. 2. (a) *Simplified model of the breast region with a cancerous lesion D under the scan probe.* (b) *Schematic of the scan probe in the (x_1, x_2)-plane.*

and

$$V_0 \int_\Gamma \Im(g_\omega) ds = \min_{v\in\mathcal{H}_{re}} \max_{h\in\mathcal{H}_{im}} \int_\Omega \left[\omega\epsilon|\nabla v|^2 + 2\sigma\nabla v\cdot\nabla h - \omega\epsilon|\nabla h|^2\right] dx, \tag{4}$$

where $\mathcal{H}_{re} := \{v \in H^1(\Omega) : v|_\Gamma = 0,\ v|_\gamma = V_0,\ \frac{\partial v}{\partial \mathbf{n}}|_{\partial\Omega\setminus(\Gamma\cup\gamma)} = 0\}$ and $\mathcal{H}_{im} := \{h \in H^1(\Omega) : h|_{\Gamma\cup\gamma} = 0,\ \frac{\partial h}{\partial \mathbf{n}}|_{\partial\Omega\setminus(\Gamma\cup\gamma)} = 0\}$.

In order to detect a lesion D underneath the scan probe Γ, we define a local region of interest under the probe plane Γ as shown in Figure 2. For simplicity, we let x_3 be the axis normal to Γ and let the center of Γ be the origin. Hence, the probe region Γ can be approximated as a two-dimensional region $\Gamma = \{(x_1, x_2, 0) : \sqrt{x_1^2 + x_2^2} < L\}$, where L is the radius of the scan probe. We set the region of interest inside the breast as a half ball $\Omega_L = \Omega \cap B_L$ shown in Figure 2, where B_L is a ball with a radius L and its center at the origin.

Remark 2.1. We summarize conductivity and permittivity values of normal and tumor tissues in the breast. Both σ and $\omega\epsilon$ have a unit of S/m and $\sigma + i\omega\epsilon = \sigma + i2\pi f\epsilon_0\epsilon_r$, where $\epsilon_0 \approx 8.854 \times 10^{-12}$ [F/m] is the permittivity of the free space and ϵ_r is a relative permittivity. Note that $\frac{\omega\epsilon_n}{\sigma_n} \le \frac{1}{50}$ for a frequency $f = \omega/2\pi \ge$50 kHz [37].

$f = \omega/2\pi$, [Hz]	σ_n, [S/m]	σ_c, [S/m]	$\omega\epsilon_n$, [S/m]	$\omega\epsilon_c$, [S/m]
≤ 500	0.03	0.2	$\ll \sigma_n$	$\ll \sigma_c$
50×10^3	0.03	0.2	5.6×10^{-4}	1.7×10^{-2}
100×10^3	0.03	0.2	2.8×10^{-4}	2.2×10^{-2}
500×10^3	0.03	0.2	1.1×10^{-3}	5.6×10^{-2}

For a successful anomaly detection, we should carefully choose two frequencies ω_1 and ω_2. In our TAS system, we choose $f_1 = \omega_1/2\pi$ and $f_2 = \omega_2/2\pi$ such that

$$50 \text{ Hz} \le f_1 \le 500 \text{ Hz} \quad \text{and} \quad 50 \text{ kHz} \le f_2 \le 500 \text{ kHz}. \tag{5}$$

We denote by $u_1 = v_1 + ih_1$ and $u_2 = v_2 + ih_2$ the complex potentials satisfying (2) at ω_1 and ω_2, respectively, and let $g_1 = g_{\omega_1}$ and $g_2 = g_{\omega_2}$. The fdTAS aims to detect D from a weighted difference between g_1 and g_2.

Now, let us investigate the connection between u_1 and u_2 and whether the frequency-difference Neumann data $g_2 - \alpha g_1$ has any information of D. Since both σ and ϵ depend on ω and x, $\sigma(x, \omega_1) \ne \sigma(x, \omega_2)$ and $\epsilon(x, \omega_1) \ne \epsilon(x, \omega_2)$. For simplicity, we denote

$$\sigma_j(x) = \sigma(x, \omega_j) \qquad \text{and} \qquad \epsilon_j(x) = \epsilon(x, \omega_j), \qquad j = 1, 2.$$

There is a cancerous lesion D inside Ω_L, and the complex conductivity $\sigma_j + i\omega_j\epsilon_j$ changes abruptly across ∂D as in the table in Remark 2.1. To distinguish them, we denote

$$(6)\qquad \sigma_j = \begin{cases} \sigma_{j,n} & \text{in } \Omega_L \setminus \overline{D}, \\ \sigma_{j,c} & \text{in } D, \end{cases} \qquad \text{and} \qquad \epsilon_j = \begin{cases} \epsilon_{j,n} & \text{in } \Omega_L \setminus \overline{D}, \\ \epsilon_{j,c} & \text{in } D. \end{cases}$$

With the use of this notation, u_1 and u_2 satisfy

$$(7) \begin{cases} \nabla \cdot ((\sigma_1 + i\omega_1\epsilon_1)\nabla u_1) = 0 \quad \text{in } \Omega, \\ u_1|_\Gamma = 0, \quad u_1|_\gamma = V_0, \\ (\sigma_1 + i\omega_1\epsilon_1)\frac{\partial u_1}{\partial \mathbf{n}}|_{\partial\Omega\setminus(\Gamma\cup\gamma)} = 0, \end{cases} \text{and} \begin{cases} \nabla \cdot ((\sigma_2 + i\omega_2\epsilon_2)\nabla u_2) = 0 \quad \text{in } \Omega, \\ u_2|_\Gamma = 0, \quad u_2|_\gamma = V_0, \\ (\sigma_2 + i\omega_2\epsilon_2)\frac{\partial u_2}{\partial \mathbf{n}}|_{\partial\Omega\setminus(\Gamma\cup\gamma)} = 0. \end{cases}$$

Remark 2.2. Due to the complicated structure of (3) and (4) for the solution u_ω, it is quite difficult to analyze the interrelation between the complex conductivity contrast $\nabla(\sigma + i\omega\epsilon)$ and the Neumann data g_ω. In [1], the authors briefly mentioned that the multifrequency TAS method can be regarded as a straightforward extension of their single-frequency TAS algorithm (Remark 2.3 in [1]). However, the simple frequency-difference data $g_2 - g_1$ on Γ may fail to extract the anomaly for more general cases of complex conductivity distributions in Ω due to the complicated structure of the solution of (2). To be precise, the use of the weighted difference is essential when the background comprises biological materials with nonnegligible frequency-dependent complex conductivity values. To explain it clearly, consider a homogeneous complex conductivity distribution in Ω where $\sigma(x,\omega)+i\omega\epsilon(x,\omega)$ depends only on ω. Due to the frequency dependency, the simple difference $g_2 - g_1$ is not zero, while $g_2 - \alpha g_1 = 0$. Hence, any reconstruction method using $g_2 - g_1$ always produces artifacts because $g_2 - g_1$ does not eliminate modeling errors. See (26) for an approximation of $g_2 - g_1$ in the presence of an anomaly D.

Remark 2.3. In this work, we do not consider effects of contact impedances along electrode-skin interfaces. For details about the contact impedance, please see [20, 36] and other publications cited therein. In TAS, we may adopt a skin preparation procedure and electrode gels to reduce contact impedances. Since we cannot expect complete removal of contact impedances, however, we need to investigate how exit currents are affected by contact impedances of a planner array of electrodes that are kept at the grounded potential. The contact impedance of each electrode leads to a voltage drop across it, and therefore the voltage underneath the electrode-skin interface layer would be slightly different than zero. In other words, when contact impedances are not negligible, the surface area in contact with Γ cannot be regarded as an equipotential surface anymore, and this will result in some changes in exit currents. Future studies are needed to estimate how the contact impedance affects the weighted difference of the Neumann data. We should also investigate experimental techniques, including choice of frequencies, to minimize their effects.

The next observation explains why we should use a weighted difference $g_2 - \alpha g_1$ instead of $g_2 - g_1$.

Observation 2.4. Denoting $\eta := \frac{\sigma_2 + i\omega_2\epsilon_2}{\sigma_1 + i\omega_1\epsilon_1}$, it follows from a direct computation that $u_2 - u_1$ satisfies

$$(8)\qquad \begin{cases} \nabla \cdot ((\sigma_1 + i\omega_1\epsilon_1)\nabla(u_2 - u_1)) = -(\sigma_1 + i\omega_1\epsilon_1)\nabla \log\eta \cdot \nabla u_2 \quad \text{in } \Omega, \\ (u_2 - u_1)|_{\Gamma\cup\gamma} = 0, \\ (\sigma_1 + i\omega_1\epsilon_1)\frac{\partial(u_2 - u_1)}{\partial \mathbf{n}}|_{\partial\Omega\setminus(\Gamma\cup\gamma)} = 0. \end{cases}$$

For the detection of D, we use the following weighted difference:

$$g_2 - \alpha g_1 = \eta\ (\sigma_1 + i\omega_1\epsilon_1)\ \mathbf{n} \cdot \nabla(u_2 - u_1) \qquad \text{on } \Gamma,$$

where $\alpha = \eta|_\Gamma$. If $\nabla \log \eta = 0$ in (8), $u_1 = u_2$ in Ω and $g_2 - \alpha g_1 = 0$ on Γ. In other words, if $\nabla \log \eta = 0$ in Ω_L, it is impossible to detect D from $g_2 - \alpha g_1 = 0$ regardless of contrasts in σ and ϵ across ∂D. Any useful information on D could be found from nonzero $g_2 - \alpha g_1$ on Γ when $|\nabla \log \eta|$ is large along ∂D.

For chosen frequencies ω_1 and ω_2, we can assume that σ and ϵ are approximately constant in the normal breast region $\Omega_L \setminus \overline{D}$ and also in the cancerous region D. Hence, if η changes abruptly across ∂D, we roughly have

$$\nabla \log \eta \approx 0 \quad \text{in } \Omega_L \setminus \bar{D} \quad \text{and} \quad |\nabla \log \eta| = \infty \quad \text{on } \partial D,$$

and therefore the term $(\sigma_1 + i\omega_1\epsilon_1)\nabla \log \eta \cdot \nabla u_2$ in (8) is supported on ∂D in the breast region Ω_L. This explains that the difference $g_2 - \alpha g_1$ on Γ can provide the information of ∂D. Take note that the inner product $\nabla \log \eta \cdot \nabla u_2$ is to be interpreted in a suitable distributional sense if the coefficients jump at ∂D.

3. Mathematical analysis for fdTAS.

3.1. Representation formula. Observation 2.4 in the previous section roughly explains how D is related to $g_2 - \alpha g_1$. In this section, the observation will be justified rigorously in a simplified model. We assume that $\sigma_{j,n}$, $\sigma_{j,c}$, $\epsilon_{j,n}$, and $\epsilon_{j,c}$ are constants. According to the table in Remark 2.1, the change of the conductivity due to the change of frequency is small, so we assume that

$$\sigma_{1,n} = \sigma_{2,n} := \sigma_n \quad \text{and} \quad \sigma_{1,c} = \sigma_{2,c} := \sigma_c. \tag{9}$$

Since the breast region of interest is relatively small compared with the entire body Ω, we may assume that Ω is the lower half space $\Omega = \mathbb{R}^3_- := \{\mathbf{x} = (x_1, x_2, x_3) \mid x_3 < 0\}$ and $\gamma = \infty$.

Suppose that v_j and h_j are H^1-solutions of the following coupled system for j =1 and 2:

$$\begin{cases} \nabla \cdot (\sigma \nabla v_j) - \nabla \cdot (\omega_j \epsilon_j \nabla h_j) = 0 & \text{in} \quad \Omega = \mathbb{R}^3_-, \\ \nabla \cdot (\omega_j \epsilon_j \nabla v_j) + \nabla \cdot (\sigma \nabla h_j) = 0 & \text{in} \quad \Omega = \mathbb{R}^3_-, \\ \quad v_j = 1 \quad \text{and} \quad h_j = 0 & \text{on} \quad \Gamma, \\ \mathbf{n} \cdot \nabla v_j = 0 \quad \text{and} \quad \mathbf{n} \cdot \nabla h_j = 0 & \text{on} \quad \partial\Omega \setminus \Gamma. \end{cases} \tag{10}$$

Let $u_j = v_j + ih_j$. Then $V_0(1 - u_j)$ can be viewed as a solution of (7) with $\Omega = \mathbb{R}^3_-$ and $\gamma = \infty$.

Let us introduce a key representation formula explaining the relationship between D and the weighted difference $g_2 - \alpha g_1$. For each $x \in \mathbb{R}^3 \backslash \Gamma$, we define

$$\Psi(x, y) = \Phi(x, y) + \Phi(x, y^+) + \varphi(x, y),$$

where $y^+ = (y_1, y_2, -y_3)$ is the reflection point of y with respect to the plane $\{y_3 = 0\}$ and $\varphi(x, \cdot)$ is the $H^1(\mathbb{R}^3 \backslash \Gamma)$-solution of the following PDE:

$$\begin{cases} \Delta_y \varphi(x, y) = 0, \quad y \in \mathbb{R}^3 \setminus \Gamma, \\ \varphi(x, y) = \frac{1}{2\pi|x-y|}, \quad y \in \Gamma, \\ \varphi(x, y) = 0 \quad \text{as } |y| \to \infty. \end{cases}$$

The following theorem explains an explicit relation between D and $\Im(g_2 - \alpha g_1)$.

THEOREM 3.1. *The imaginary part of the weighted difference $g_2 - \alpha g_1$ satisfies the following formula:*

$$\frac{1}{2\sigma_n}\Im(g_2-\alpha g_1)(x) = \int_D \nabla_y \frac{\partial\Phi(x,y)}{\partial x_3}\cdot\Theta(y)dy + \frac{\partial}{\partial x_3}\int_{\partial\Omega\setminus\Gamma}\frac{\partial\Phi(x,y)}{\partial y_3}\left[\int_D \nabla_z\Psi(y,z)\cdot\Theta(z)dz\right]ds,\ x\in\Gamma, \tag{11}$$

where

$$\Theta(y) = \frac{\sigma_n-\sigma_c}{\sigma_n}\nabla(h_2-h_1)(y) + \frac{\omega_2(\epsilon_{2,n}-\epsilon_{2,c})}{\sigma_n}\nabla(v_2-v_1)(y) - \Im(\beta\nabla u_1(y))$$

and

$$\beta = \frac{i}{1+i\frac{\omega_1\epsilon_{1,n}}{\sigma_n}} \cdot\left[\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\left(\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\right) - \frac{\omega_1\epsilon_{1,n}}{\sigma_n}\left(\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\sigma_c}{\sigma_n}\right) - i\frac{\omega_1\omega_2\epsilon_{1,n}\epsilon_{2,n}}{\sigma_n^2}\left(\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\epsilon_{2,c}}{\epsilon_{2,n}}\right)\right].$$

Proof. Due to Green's identity, $(\sigma_n+i\omega_2\epsilon_{2,n})(u_2-u_1)$ has integral representation: for each $x\in\Omega$,

$$\begin{aligned}&(\sigma_n+i\omega_2\epsilon_{2,n})(u_2-u_1)(x)\\ &= (\sigma_n+i\omega_2\epsilon_{2,n})\int_{\partial\Omega\setminus\Gamma}\frac{\partial\Phi(x,y)}{\partial\mathbf{n}}(u_2-u_1)(y)ds + \int_\Gamma \Phi(x,y)(g_2-\alpha g_1)(y)\ ds\\ &+ \int_{\partial D}\Phi(x,y)(\sigma_n+i\omega_2\epsilon_{2,n})\left(\frac{\partial(u_2-u_1)}{\partial\mathbf{n}}|_+ - \frac{\partial(u_2-u_1)}{\partial\mathbf{n}}|_-\right)(y)\ ds.\end{aligned} \tag{12}$$

Here, we denote $\frac{\partial u_j}{\partial\mathbf{n}}|_\pm = \mathbf{n}\cdot\nabla u_j^\pm|_{\partial D}$, where $u_j^+ = u_j|_{\Omega\setminus\bar D}$ and $u_j^- = u_j|_D$. From the transmission condition,

$$(\sigma_n+i\omega_j\epsilon_{j,n})\frac{\partial u_j}{\partial\mathbf{n}}|_+ = (\sigma_c+i\omega_j\epsilon_{j,c})\frac{\partial u_j}{\partial\mathbf{n}}|_-,\quad j=1,2 \qquad \text{on } \partial D.$$

It follows that

$$\frac{\sigma_n+i\omega_2\epsilon_{2,n}}{\sigma_n}\frac{\partial(u_2-u_1)}{\partial\mathbf{n}}|_+ = \frac{\sigma_c+i\omega_2\epsilon_{2,c}}{\sigma_n}\frac{\partial(u_2-u_1)}{\partial\mathbf{n}}|_- + \beta\frac{\partial u_1}{\partial\mathbf{n}}|_- \quad \text{on } \partial D. \tag{13}$$

Putting (13) into (12) and then applying $-\frac{\partial}{\partial x_3}$ to both sides of (12) yield for each $x\in\Gamma$

$$\frac{1}{2\sigma_n}\Im(g_2-\alpha g_1)(x) = \int_D \nabla_y\frac{\partial\Phi(x,y)}{\partial x_3}\cdot\left[\frac{\sigma_n-\sigma_c}{\sigma_n}\nabla(h_2-h_1) + \frac{\omega_2(\epsilon_{2,n}-\epsilon_{2,c})}{\sigma_n}\nabla(v_2-v_1) - \Im(\beta\nabla u_1)\right]dy + \Xi(x), \tag{14}$$

where

$$\Xi(x) = -\frac{\partial}{\partial x_3}\int_{\partial\Omega\setminus\Gamma}\frac{\partial\Phi(x,y)}{\partial y_3}\left((h_2-h_1)(y) + \frac{\omega_2\epsilon_{2,n}}{\sigma_n}(v_2-v_1)(y)\right)ds.$$

From the definition of $\Psi(x,y)$, it is easy to see that $\Psi(x,y)$ satisfies

$$\begin{cases} \triangle_x \Psi(x,y) = \delta(x-y), \quad x,\ y \in \mathbb{R}^3_-, \\ \Psi(x,y) = 0, \quad x \in \Gamma,\ y \in \mathbb{R}^3_-, \\ \frac{\partial \Psi(x,y)}{\partial x_3} = 0, \quad x \in \partial\Omega \setminus \Gamma,\ y \in \mathbb{R}^3_-, \\ \Psi(x,y) = 0 \quad \text{as } |x-y| \to \infty. \end{cases}$$

In order to relate $\Xi(x)$ with D, we repeat the argument in (12) with Φ replaced by Ψ:

$$(h_2 - h_1)(y) + \frac{\omega_2 \epsilon_{2,n}}{\sigma_n}(v_2 - v_1)(y) = \frac{1}{\sigma_n} \Im \left\{ (\sigma_n + i\omega_2\epsilon_{2,n})(u_2 - u_1)(y) \right\}$$

$$= \Im \left\{ \int_{\partial D} \Psi(y,z) \frac{\sigma_n + i\omega_2\epsilon_{2,n}}{\sigma_n} \left(\frac{\partial(u_2 - u_1)}{\partial \mathbf{n}}|_+ - \frac{\partial(u_2 - u_1)}{\partial \mathbf{n}}|_- \right)(z)\ ds \right\},\ y \in \partial\Omega \setminus \Gamma.$$

The above identity and the jump condition (13) lead to

$$(h_2 - h_1)(y) + \frac{\omega_2 \epsilon_{2,n}}{\sigma_n}(v_2 - v_1)(y) = -\int_D \nabla_z \Psi(y,z) \cdot \Theta(z) dz, \quad y \in \partial\Omega \setminus \Gamma.$$

This completes the proof. □

Now, let us derive a constructive formula extracting D from the representation formula (11) under some reasonable assumptions. We assume that

$$\text{(15)} \qquad \bar{D} \subset \Omega_{L/2}, \qquad D = B_\delta(\xi), \qquad \text{and} \qquad \delta \le \operatorname{dist}(D,\Gamma) \le C_1\delta,$$

where C_1 is a positive constant, B_δ is a ball with the radius δ and the center ξ, and $\frac{\delta}{L} \le \frac{1}{10}$. Suppose we choose $\frac{\omega_1}{2\pi} \approx 50$ Hz and $\frac{\omega_2}{2\pi} \approx 100$ kHz. Then the experimental data in Remark 2.1 shows $\frac{\omega_2\epsilon_{2,n}}{\sigma_n} \approx \frac{1}{100}$ and $\frac{\omega_1\epsilon_{1,n}}{\sigma_n} \le \frac{1}{10000}$. Hence, in practice, we can view

$$\text{(16)} \qquad \frac{\omega_1\epsilon_{1,n}}{\sigma_n} \approx 0, \quad (\delta/L)^3 \approx 0, \quad \left(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\right)^2 \approx 0.$$

Based on the experimental data in Remark 2.1, we assume that

$$\text{(17)} \qquad \max\left\{\frac{\epsilon_{j,n}}{\epsilon_{j,c}}, \frac{\sigma_n}{\sigma_c}\right\} \le \kappa_1, \quad \frac{\omega_2\epsilon_{2,n}}{\sigma_n} \le \kappa_2\frac{\sigma_n}{\sigma_c}, \quad \frac{\sigma_c}{\sigma_n} \le \kappa_3,$$

where κ_1 and κ_2 are positive constants less than $\frac{1}{2}$ and κ_3 is a positive constant less than 10. Taking advantage of these, we can simplify the representation formula (11).

THEOREM 3.2. *Under the assumptions* (15) *and* (17), *the imaginary part of the weighted frequency difference* $g_2 - \alpha g_1$ *can be expressed as*

$$\text{(18)} \quad \frac{1}{2\sigma_n}\Im\left(g_2 - \alpha g_1\right)(x) = \int_D \frac{\partial}{\partial x_3}\frac{(x-y)\cdot\tilde{\Theta}(y)}{4\pi|x-y|^3}dy + \mathit{Error}(x), \quad x \in \Gamma_{L/2},$$

where

$$\tilde{\Theta} = \frac{\sigma_n - \sigma_c}{\sigma_n}\nabla h_2 - \frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big(\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n}\Big)\nabla v_1$$

and the error term $Error(x)$ is estimated by

$$|Error(x)| \leq \Bigg[\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\mathcal{P}_1\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\frac{\delta^3}{L^3}$$

$$+\left(\frac{\omega_1\epsilon_{1,n}}{\sigma_n}\mathcal{P}_1\Big(\Big|\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)+\Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\right)\frac{\delta^3}{|x-\xi|^3}\Bigg].$$

Here, $\mathcal{P}_n(\lambda)$ is a polynomial function of order n such that $\mathcal{P}_n(0)=0$ and its coefficients depend only on $\kappa_j, j=1,2,3$.

Proof. From the transmission conditions of u_ω across ∂D, we have

$$\frac{\partial h_\omega}{\partial \mathbf{n}}\Big|_+ - \frac{\sigma_c}{\sigma_n}\frac{\partial h_\omega}{\partial \mathbf{n}}\Big|_- = \frac{\frac{\omega\epsilon_n}{\sigma_n}}{1+(\frac{\omega\epsilon_n}{\sigma_n})^2}\Big(\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big)\frac{\partial v_\omega}{\partial \mathbf{n}}\Big|_- + \frac{(\frac{\omega\epsilon_n}{\sigma_n})^2}{1+(\frac{\omega\epsilon_n}{\sigma_n})^2}\Big(\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big)\frac{\partial h_\omega}{\partial \mathbf{n}}\Big|_-.$$

Since h_ω satisfies the mixed boundary condition with $h_\omega|_\Gamma = 0$ and $\frac{\partial h_\omega}{\partial \mathbf{n}}|_{\partial\Omega\setminus\Gamma}=0$, we have the following estimate:

$$\int_\Omega \Big(\chi_{\Omega\setminus\bar{D}}+\frac{\sigma_c}{\sigma_n}\chi_D\Big)|\nabla h_\omega|^2 \leq \frac{\omega\epsilon_n}{\sigma_n}\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\|\nabla v_\omega\|_{L^2(D)}\|\nabla h_\omega\|_{L^2(D)} + \Big(\frac{\omega\epsilon_n}{\sigma_n}\Big)^2\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\|\nabla h_\omega\|^2_{L^2(D)}. \tag{19}$$

This gives

$$\|\nabla h_\omega\|_{L^2(D)} \leq \left(\frac{\sigma_c}{\sigma_n}-\Big(\frac{\omega\epsilon_n}{\sigma_n}\Big)^2\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\right)^{-1}\Big(\frac{\omega\epsilon_n}{\sigma_n}\Big)\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\|\nabla v_\omega\|_{L^2(D)}. \tag{20}$$

Since $\|\nabla v_\omega\|_{L^2(D)} \leq C\sqrt{|D|}$, where C depends only on κ_3, we obtain

$$\|\nabla h_\omega\|_{L^2(D)} \leq \Big(\frac{\omega\epsilon_n}{\sigma_n}\Big)\mathcal{P}_1\Big(\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\sqrt{|D|}. \tag{21}$$

We also use the jump condition for v_ω:

$$\frac{\partial v_\omega}{\partial \mathbf{n}}\Big|_+ - \frac{\sigma_c}{\sigma_n}\frac{\partial v_\omega}{\partial \mathbf{n}}\Big|_- = \frac{(\frac{\omega\epsilon_n}{\sigma_n})^2}{1+(\frac{\omega\epsilon_n}{\sigma_n})^2}\Big(\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big)\frac{\partial v_\omega}{\partial \mathbf{n}}\Big|_- + \frac{\frac{\omega\epsilon_n}{\sigma_n}}{1+(\frac{\omega\epsilon_n}{\sigma_n})^2}\Big(\frac{\sigma_c}{\sigma_n}-\frac{\epsilon_c}{\epsilon_n}\Big)\frac{\partial h_\omega}{\partial \mathbf{n}}\Big|_-.$$

Applying the same process as in (19), we obtain

$$\|\nabla v_\omega - \nabla u_0\|_{L^2(D)} \leq \left(\frac{\omega\epsilon_n}{\sigma_n}\right)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_c}{\epsilon_n}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\sqrt{|D|}, \tag{22}$$

where $u_0 = u_\omega$ with $\omega = 0$. From (22), we get

$$\|\nabla v_2 - \nabla v_1\|_{L^2(D)} \leq \|\nabla v_2 - \nabla u_0\|_{L^2(D)} + \|\nabla v_1 - \nabla u_0\|_{L^2(D)}$$

$$\leq \left(\Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)+\Big(\frac{\omega_1\epsilon_{1,n}}{\sigma_n}\Big)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\right)\sqrt{|D|}. \tag{23}$$

Hence, it follows from (20) and (23) that

$$\|\Theta - \tilde{\Theta}\|_{L^2(D)} \leq \left(\frac{\omega_1\epsilon_{1,n}}{\sigma_n}\mathcal{P}_1\Big(\Big|\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)+\Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\right)\sqrt{|D|}.$$

From the Schwarz inequality,

$$\left|\int_D \frac{\partial}{\partial x_3}\frac{(x-y)}{4\pi|x-y|^4}\cdot\Big(\Theta(y)-\tilde{\Theta}(y)\Big)dy\right| \le \frac{C\sqrt{|D|}}{|x-\xi|^3}\|\Theta-\tilde{\Theta}\|_{L^2(D)}$$

$$\le \left(\frac{\omega_1\epsilon_{1,n}}{\sigma_n}\mathcal{P}_1\Big(\Big|\frac{\epsilon_{1,c}}{\epsilon_{1,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)+\Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)^2\mathcal{P}_2\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\Big|\Big)\right)\frac{\delta^3}{|x-\xi|^3}.$$

Now, it remains to study the last term in (11). Using the Schwarz inequality, it is easy to see that

$$\left|\frac{\partial}{\partial x_3}\int_{\partial\Omega\setminus\Gamma}\frac{\partial\Phi(x,y)}{\partial y_3}\int_D \nabla_z\Psi(y,z)\cdot\Theta(z)dz\right| \le \frac{\delta^{\frac{3}{2}}}{L^3}\|\Theta\|_{L^2(D)}. \tag{24}$$

This completes the proof. □

Remark 3.3. According to Theorem 3.2, (21), and (23),

$$\frac{1}{2\sigma_n}\Im\left(g_2-\alpha g_1\right)=0 \quad\text{when}\quad \left|\frac{\epsilon_{j,c}}{\epsilon_{j,n}}-\frac{\sigma_c}{\sigma_n}\right|=0,\quad j=1,2.$$

Hence, even if $\epsilon_{2,c}$ and $\epsilon_{1,c}$ are quite different, we cannot extract any information of D when $|\frac{\epsilon_{j,c}}{\epsilon_{j,n}}-\frac{\sigma_c}{\sigma_n}|=0, j=1,2$. On the other hand, even if $\epsilon_{2,c}=\epsilon_{1,c}$, we can extract the information of D whenever $|\frac{\epsilon_{j,c}}{\epsilon_{j,n}}-\frac{\sigma_c}{\sigma_n}|\neq 0, j=1,2$.

Remark 3.4. Based on (18), we can derive the following simple approximate formula for the reconstruction of D:

$$\begin{aligned}\frac{1}{2\sigma_n}\Im\left(g_2-\alpha g_1\right)(x) \approx{}& \frac{\omega_2\epsilon_{2,n}}{\sigma_n}\frac{(3\sigma_n)^2}{(2\sigma_n+\sigma_c)^2}\left(\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\right)\partial_{x_3}U(\xi)\\ &\times|D|\frac{2\xi_3^2-(x_1-\xi_1)^2-(x_2-\xi_2)^2}{4\pi|x-\xi|^5},\quad x\in\Gamma_{L/2},\end{aligned} \tag{25}$$

where U is the solution of (10) in the absence of anomaly at $\omega=0$. Note that the difference g_2-g_1 can be approximated by

$$\begin{aligned}\frac{1}{2\sigma_n}\left(g_2-g_1\right)(x) \approx{}& i\frac{\omega_2\epsilon_{2,n}}{2\sigma_n^2}g_1(x)+i\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\frac{(3\sigma_n)^2}{(2\sigma_n+\sigma_c)^2}\left(\frac{\epsilon_{2,c}}{\epsilon_{2,n}}-\frac{\sigma_c}{\sigma_n}\right)\partial_{x_3}U(\xi)\\ &\times|D|\frac{2\xi_3^2-(x_1-\xi_1)^2-(x_2-\xi_2)^2}{4\pi|x-\xi|^5},\quad x\in\Gamma_{L/2},\end{aligned} \tag{26}$$

and therefore any detection algorithm using the above approximation will be disturbed by the term $\frac{\omega_2\epsilon_{2,n}}{2\sigma_n^2}g_1$.

We will prove the approximation (25) roughly. From [1], we have

$$\Big|\nabla U(y)-\partial_{x_3}U(\xi)\mathbf{e}_3\Big| \le C\frac{\delta}{L}\sqrt{|D|},\quad y\in D, \tag{27}$$

where $\mathbf{e}_3=(0,0,1)$. Now let V be the H^1-solution of the following PDE:

$$\begin{cases}\Delta V=0 \quad \text{in } \Omega\setminus\partial D,\\ \frac{\sigma_n}{\sigma_c}\frac{\partial V}{\partial\mathbf{n}}|_+-\frac{\partial V}{\partial\mathbf{n}}|_- = -\mathbf{n}\cdot\mathbf{e}_3 \quad \text{on } \partial D,\\ V|_\Gamma=0 \quad\text{and}\quad \frac{\partial V}{\partial\mathbf{n}}|_{\partial\Omega\setminus\Gamma}=0.\end{cases} \tag{28}$$

Using the same process as in (19) and (27), we obtain

$$\left\|\nabla u_0 - \nabla U(y) - \partial_{x_3} U(\xi)\Big(\frac{\sigma_n}{\sigma_c} - 1\Big)\nabla V\right\|_{L^2(D)} \le C\frac{\delta}{L}\sqrt{|D|}, \tag{29}$$

where C depends only on κ_3. From (22), (27), and (29), we have

$$\begin{aligned}&\left\|\nabla v_\omega - \partial_{x_3} U(\xi)\Big(\mathbf{e}_3 + \Big(\frac{\sigma_n}{\sigma_c} - 1\Big)\nabla V\Big)\right\|_{L^2(D)} \\ &\qquad \le \left(\Big(\frac{\omega\epsilon_n}{\sigma_n}\Big)^2 \mathcal{P}_2\Big(\Big|\frac{\epsilon_c}{\epsilon_n} - \frac{\sigma_c}{\sigma_n}\Big|\Big) + C\frac{\delta}{L}\right)\sqrt{|D|}.\end{aligned} \tag{30}$$

Applying the same process as in (19) and the identities (20) and (29), we obtain

$$\begin{aligned}&\left\|\nabla h_2 - \frac{\omega_2\epsilon_{2,n}}{\sigma_n}\frac{\sigma_n}{\sigma_c}\Big(\frac{\sigma_c}{\sigma_n} - \frac{\epsilon_{2,c}}{\epsilon_{2,n}}\Big)\partial_{x_3}U(\xi)\Big(1 + \Big(\frac{\sigma_n}{\sigma_c} - 1\Big)\partial_{x_3}V(\xi)\Big)\nabla V\right\|_{L^2(D)} \\ &\le \Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)\mathcal{P}_1\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n}\Big|\Big)\left\|\nabla v_2 - \partial_{x_3}U(\xi)\Big(\mathbf{e}_3 + \Big(\frac{\sigma_n}{\sigma_c} - 1\Big)\partial_{x_3}V(\xi)\mathbf{e}_3\Big)\right\|_{L^2(D)}.\end{aligned}$$

From (30), we get

$$\begin{aligned}&\left\|\nabla h_2 - \frac{\omega_2\epsilon_{2,n}}{\sigma_n}\frac{\sigma_n}{\sigma_c}\Big(\frac{\sigma_c}{\sigma_n} - \frac{\epsilon_{2,c}}{\epsilon_{2,n}}\Big)\partial_{x_3}U(\xi)\Big(\mathbf{e}_3 + \Big(\frac{\sigma_n}{\sigma_c} - 1\Big)\nabla V(\xi)\Big)\nabla V\right\|_{L^2(D)} \\ &\qquad \le \left(\Big(\frac{\omega_2\epsilon_{2,n}}{\sigma_n}\Big)^3\mathcal{P}_3\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n}\Big|\Big) + \frac{\omega_2\epsilon_{2,n}}{\sigma_n}\mathcal{P}_1\Big(\Big|\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n}\Big|\Big)\frac{\delta}{L}\right)\sqrt{|D|}.\end{aligned} \tag{31}$$

From [1], ∇V in D is approximated by

$$\nabla V|_D \approx \frac{\sigma_c}{2\sigma_n + \sigma_c}\Big(1 - \frac{r^3}{16\pi|\xi_3|^3}\Big)\mathbf{e}_3 \approx \frac{\sigma_c}{2\sigma_n + \sigma_c}\mathbf{e}_3. \tag{32}$$

The approximation (25) follows immediately from Theorem 3.2, (30), (31), and (32).

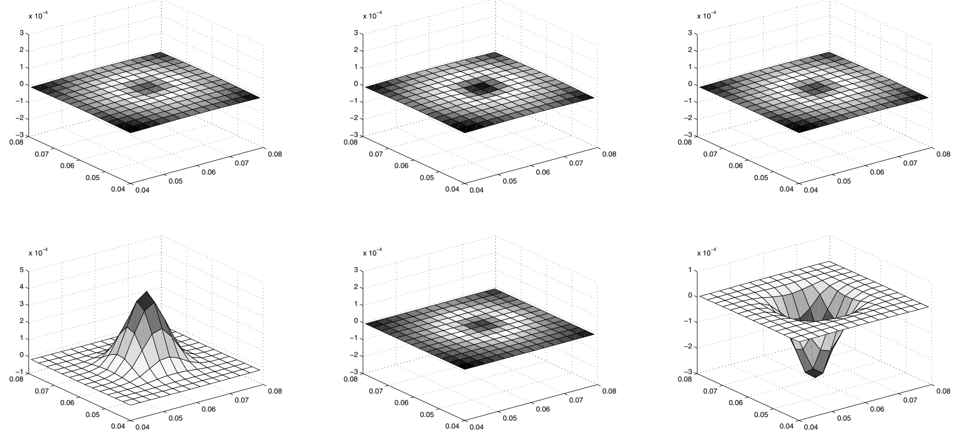
Remark 3.5. Our reconstruction algorithm is based on the approximation formula (25). In practice, we may not have a priori knowledge of the background conductivities. In that case, α is unknown. But α can be evaluated approximately by the ratio of the measured Neumann data as follows:

$$\alpha \;=\; \frac{\int_\Gamma g_{\omega_2} ds}{\int_\Gamma g_{\omega_1} ds} \;+\; \frac{\int_D \left((1-\alpha)\sigma_c + i(\omega_2\epsilon_{2,c} - \alpha\omega_1\epsilon_{1,c})\right)\nabla u_2 \cdot \nabla u_1 dx}{\int_\Omega (\sigma + i\omega_1\epsilon_1)|\nabla u_1|^2 dx}. \tag{33}$$

Hence, we may choose $\alpha \approx \frac{\int_\Gamma g_{\omega_2} ds}{\int_\Gamma g_{\omega_1} ds}$.

We can prove the identity (33) for a bounded domain Ω. Using $u_1|_\gamma = u_2|_\gamma = V_0$, we have

$$\begin{aligned}\textstyle\int_\Gamma (g_2 - \alpha g_1) ds \quad &= -\textstyle\int_\gamma (g_2 - \alpha g_1)\, ds = -\frac{1}{V_0}\int_\gamma (g_2 u_1 - \alpha g_1 u_1) ds \\ &= -\tfrac{1}{V_0}\textstyle\int_{\partial\Omega}(g_2 - \alpha g_1) u_1 ds \\ &= \tfrac{1}{V_0}\textstyle\int_\Omega \left((\sigma + i\omega_1\epsilon_2)\nabla u_2 - \alpha(\sigma + i\omega_1\epsilon_1)\nabla u_1\right)\cdot\nabla u_1 dx \\ &= \tfrac{1}{V_0}\textstyle\int_\Omega \Big[\alpha(\sigma + i\omega_1\epsilon_1)\left(\nabla u_2 - \nabla u_1\right)\cdot\nabla u_1 \\ &\qquad\qquad + \left((1-\alpha)\sigma + i(\omega_2\epsilon_2 - \alpha\omega_1\epsilon_1)\right)\nabla u_2 \cdot \nabla u_1\Big] dx \\ &= \tfrac{1}{V_0}\textstyle\int_D \left((1-\alpha)\sigma_c + i(\omega_2\epsilon_{2,c} - \alpha\omega_1\epsilon_{1,c})\right)\nabla u_2 \cdot \nabla u_1 dx.\end{aligned}$$



(a) (b) (c)

FIG. 3. *Frequency-difference trans-admittance map: Real and imaginary parts of* $g_2 - \alpha g_1$ *with three different values of* $\omega_2 \epsilon_{2,c}$ *such that* μ *is* (a) *positive,* (b) *zero, and* (c) *negative. All the plots are shown in the region* $[\,0.04,\, 0.08\,] \times [\,0.04,\, 0.08\,]$ *m*^{3} *by using the same scale of* 10^{-4}.

The identity (33) follows from the fact that

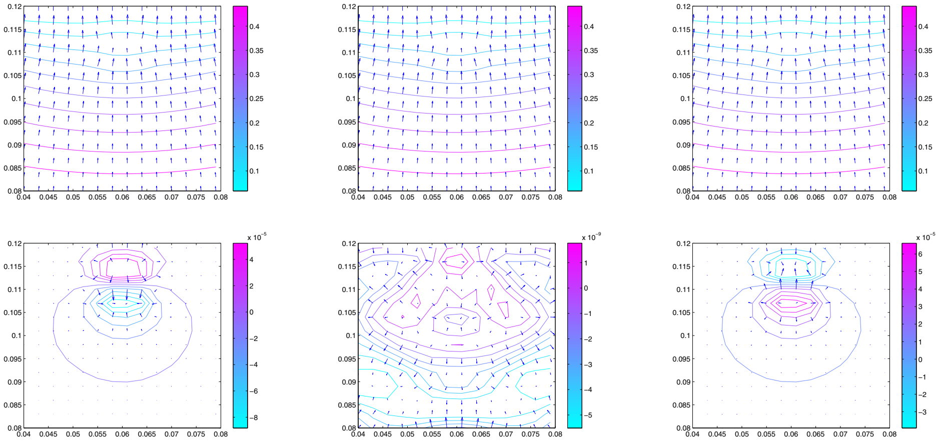
$$V_0 \int_\Gamma g_{\omega_1} ds = \int_\Omega (\sigma + i\omega_1 \epsilon_1) |\nabla u_1|^2 dx.$$

3.2. Numerical simulations. In fdTAS [32], we use a weighted frequency difference of Neumann data $g_2 - \alpha g_1$ instead of the simple difference $g_2 - g_1$. Theorems 3.1 and 3.2 show that the weight α and the difference $(\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n})$ are important factors in detecting anomaly D.

In order to test the observations in Theorem 3.2 and Remark 3.3, we consider a cubic model $\Omega := [\,0,\, 0.12\,] \times [\,0,\, 0.12\,] \times [\,0,\, 0.12\,]$ m^3 with the probe region $\Gamma := \{(x, y, 0.12) : \sqrt{x^2 + y^2} < 0.03\}$ and the reference electrode $\gamma := \{(x, y, z) \in \Omega : z = 0\}$. We assume that $\Omega \setminus D$ and D are homogeneous with frequency-independent conductivity values $\sigma_n = 0.03$ S/m and $\sigma_c = 0.2$ S/m. For permittivity values, we set $\omega_1 \epsilon_{1,n} = \omega_1 \epsilon_{1,c} = 0$ and $\omega_2 \epsilon_{2,n} = 3 \times 10^{-4}$ S/m. Numerical simulations are performed for a cube-shaped anomaly D centered at $(0.06, 0.06, 0.12 - 0.009)$ in meters with its side length of 0.006 m.

Figure 3 shows the images of $g_2 - \alpha g_1$ with three different values of $\omega_2 \epsilon_{2,c}$ that are chosen so that the corresponding $\mu = (\frac{\epsilon_{2,c}}{\epsilon_{2,n}} - \frac{\sigma_c}{\sigma_n})$ is positive, zero, or negative, respectively. This setup allows us to observe that $g_2 - \alpha g_1$ is influenced by μ and there is an interesting relation between them. As we discussed in Remark 3.3, $\mu = 0$ implies $g_2 - \alpha g_1$ providing no information on D even if ϵ_c changes a lot with respect to frequency. On the other hand, even if the permittivities ϵ_n and ϵ_c do not change with frequency, we can extract information on the anomaly from $g_2 - \alpha g_1$ as far as $\mu \neq 0$.

Figure 4 shows the vector fields of complex potential ∇u_2 corresponding to three different values of $\omega_2 \epsilon_{2,c}$ as before. In the plots, solid lines are equipotential lines of u_2, and arrows indicate the direction and magnitude of electric field $-\nabla u_2$. Figure 4



(a) (b) (c)

FIG. 4. *Equipotential lines and electric field streamlines in the slice* $\{(0.06, y, z) : 0.04 < y < 0.08, 0.08 < z < 0.12\}$*: Real and imaginary parts of the complex potential* u_2 *with three different values of* $\omega_2\epsilon_{2,c}$ *as above. Imaginary part plots are individually scaled as* (a) 10^{-5}, (b) 10^{-9}, (c) 10^{-5} *and real part plots are shown by using the same scale.*

illustrates that the electric field direction of the imaginary part changes as the sign of μ changes. We believe that the nonzero vector field is due to computational errors when $\mu = 0$.

REFERENCES

[1] H. AMMARI, O. KWON, J. K. SEO, AND E. J. WOO, *T-Scan electrical impedance imaging system for anomaly detection*, SIAM J. Appl. Math., 65 (2004), pp. 252–266.

[2] H. AMMARI, S. MOSKOW, AND M. VOGELIUS, *Boundary integral formulae for the reconstruction of electric and electromagnetic inhomogeneities of small volume*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 49–66.

[3] H. AMMARI AND J. K. SEO, *An accurate formula for the reconstruction of conductivity inhomogeneity*, Adv. in Appl. Math., 30 (2003), pp. 679–705.

[4] M. ASSENHEIMER, O. LAVER-MOSKOVITZ, D. MALONEK, D. MANOR, U. NAHLIEL, R. NITZAN, AND A. SAAD, *The T-Scan technology: Electrical impedance as a diagnostic tool for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 1–8.

[5] L. BORCEA, *EIT electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[6] M. BRÜHL AND M. HANKE, *Numerical implementation of two non-iterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[7] K. BRYAN, *Numerical recovery of certain discontinuous electrical conductivities*, Inverse Problems, 7 (1991), pp. 827–840.

[8] D. J. CEDIO-FENGYA, S. MOSKOW, AND M. VOGELIUS, *Identification of conductivity imperfections of small parameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[9] M. CHENEY, D. ISAACSON, AND J. C. NEWELL, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[10] V. CHEREPENIN, A. KARPOV, A. KORJENEVSKY, V. KORNIENKO, Y. KULTIASOV, M. OCHAPKIN, O. TROCHANOVA, AND J. MEISTER, *Three-dimensional EIT imaging of breast tissues: System design and clinical testing*, IEEE Trans. Med. Imag., 21 (2002), pp. 662–667.

[11] V. CHEREPENIN, A. KARPOV, A. KORJENEVSKY, V. KORNIENKO, A. MAZALETSKAYA, D. MAZOUROV, AND J. MEISTER, *A 3D electrical impedance tomography (EIT) system for breast cancer detection*, Physiol. Meas., 22 (2001), pp. 9–18.

[12] A. V. Cherkaev and L. V. Gibiansky, *Variational principles for complex conductivity, viscoelasticity, and similar problems in media with complex moduli*, J. Math. Phys., 35 (1994), pp. 127–145.

[13] M. H. Choi, T. J. Kao, D. Isaacson, G. J. Saulnier, and J. C. Newell, *Simplified model of a mammography geometry for breast cancer imaging with electrical impedance tomography*, in Proceedings of the 26th IEEE-EMBS Conference, San Francisco, CA, 2004, pp. 1310–1313.

[14] R. D. Cook, G. J. Saulnier, D. G. Gisser, J. G. Goble, J. C. Newell, and D. Isaacson, *ACT*3*: A high-speed, high-precision electrical impedance tomography*, IEEE Trans. Biomed. Eng., 41 (1994), pp. 713–722.

[15] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[16] S. Franco, *Design with Operational Amplifiers and Analog Integrated Circuits*, 3rd ed., McGraw-Hill, New York, 2002.

[17] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[18] A. Hartov, N. Soni, and R. Halter, *Breast cancer screening with electrical impedance tomography*, in Electrical Impedance Tomography: Methods, History and Applications, D. S. Holder, ed., IOP Publishing, Bristol, UK, 2005, pp. 167–185.

[19] R. P. Henderson and J. G. Webster, *An impedance camera for spatially specific measurements of the thorax*, IEEE Trans. Biomed. Eng., 25 (1978), pp. 250–254.

[20] N. Hyvönen, *Complete electrode model of electric impedance tomography: Approximation properties and characterization of inclusions*, SIAM J. Appl. Math., 64 (2004), pp. 902–931.

[21] M. Ikehata, *On reconstruction in the inverse conductivity problem with one measurement*, Inverse Problems, 16 (2000), pp. 785–793.

[22] J. Jossinet and M. Schmitt, *A review of parameters for the bioelectrical characterization of breast tissue*, in Electrical Bioimpedance Methods, Ann. New York Acad. Sci. 873, New York Academy of the Sciences, New York, 1999, pp. 30–41.

[23] H. Kang and J. K. Seo, *Layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[24] T. Kao, J. C. Newell, G. J. Saulinier, and D. Isaacson, *Distinguishability of inhomogeneities using planar electrode arrays and different patterns of applied excitation*, Physiol. Meas., 24 (2003), pp. 403–411.

[25] T. E. Kerner, K. D. Paulsen, A. Hartov, S. K. Soho, and S. P. Poplack, *Electrical impedance spectroscopy of the breast: Clinical imaging results in* 26 *subjects*, IEEE Trans. Med. Imag., 21 (2002), pp. 638–645.

[26] O. Kwon and J. K. Seo, *Total size estimation and identification of multiple anomalies in the inverse conductivity problem*, Inverse Problems, 17 (2001), pp. 59–75.

[27] O. Kwon, J. K. Seo, and J. R. Yoon, *A real-time algorithm for the location search of discontinuous conductivites with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[28] O. Kwon, J. R. Yoon, J. K. Seo, E. J. Woo, and Y. G. Cho, *Estimation of anomaly location and size using electrical impedance tomography*, IEEE Trans. Biomed. Eng., 50 (2003), pp. 89–96.

[29] J. L. Larson-Wiseman, *Early Breast Cancer Detection Utilizing Clustered Electrode Arrays in Impedance Imaging*, Ph.D. Thesis, RPI, Troy, NY, 1998.

[30] N. Liu, G. J. Saulnier, J. C. Newell, D. Isaacson, and T. J. Kao, *ACT*4*: A high-precision, multi-frequency electrical impedance tomography*, in Proceedings of the Conference on Biomedical Applications of Electrical Impedance Tomography, University College London, London, 2005.

[31] J. L. Mueller, D. Isaacson, and J. C. Newell, *A reconstruction algorithm for electrical impedance tomography data collected on rectangular electrode arrays*, IEEE Trans. Biomed. Eng., 46 (1999), pp. 1379–1386.

[32] T. I. Oh, J. Lee, J. K. Seo, S. W. Kim, and E. J. Woo, *Feasibility of breast cancer lesion detection using multi-frequency trans-admittance scanner (TAS) with* 10*Hz to* 500*kHz bandwidth*, Physiol. Meas., 28 (2007), pp. S71–S84.

[33] B. Scholz, *Towards virtual electrical breast biopsy: Space-frequency MUSIC for trans-admittance data*, IEEE Trans. Med. Imag., 21 (2002), pp. 588–595.

[34] J. K. Seo, O. Kwon, H. Ammari, and E. J. Woo, *Mathematical framework and anomaly estimation algorithm for breast cancer detection: Electrical impedance technique using TS2000 configuration*, IEEE Trans. Biomed. Eng., 51 (2004), pp. 1898–1906.

[35] J. E. Silva, J. P. Marques, and J. Jossinet, *Classification of breast tissue by electrical impedance spectroscopy*, Med. Biol. Eng. Comput., 38 (2000), pp. 26–30.

[36] E. Somersalo, M. Cheney, and D. Isaacson, *Existence and uniqueness for electrode models for electric current computed tomography*, SIAM J. Appl. Math., 52 (1992), pp. 1023–1040.

[37] A. J. Surowiec, S. S. Stuchly, J. R. Barr, and A. Swarup, *Dielectric properties of breast carcinoma and the surrounding tissues*, IEEE Trans. Biomed. Eng., 35 (1988), pp. 257–263.

[38] A. J. Wilson, P. Milnes, A. R. Waterworth, R. H. Smallwood, and B. H. Brown, *Mk3.5: A modular, multi-frequency successor to the Mk3a EIS/EIT system*, Physiol. Meas., 22 (2001), pp. 49–54.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 37–63

ASYMPTOTIC AND NUMERICAL TECHNIQUES FOR RESONANCES OF THIN PHOTONIC STRUCTURES[*]

J. GOPALAKRISHNAN[†], S. MOSKOW[‡], AND F. SANTOSA[§]

Abstract. We consider the problem of calculating resonance frequencies and radiative losses of an optical resonator. The optical resonator is in the form of a thin membrane with variable dielectric properties. This work provides two very different approaches for doing such calculations. The first is an asymptotic method which exploits the small thickness and high index of the membrane. We derive a limiting resonance problem as the thickness goes to zero, and for the case of a simple resonance, find a first order correction. The limiting problem and the correction are in one less space dimension, which can make the approach very efficient. Convergence estimates are proved for the asymptotics. The second approach, based on the finite element method with a truncated perfectly matched layer, is not restricted to thin structures. We demonstrate the use of these methods in numerical calculations which further illustrate their differences. The asymptotic method finds resonance by solving a dense, but small, nonlinear eigenvalue problem, whereas the finite element method yields a large but linear and sparse generalized eigenvalue problem. Both methods reproduce a localized defect mode found previously by finite difference time domain methods.

Key words. photonic band gap structure, time harmonic wave equation, thin membrane structure, resonance phenomena, nonlinear eigenvalue, asymptotic analysis, finite element method, FEM, perfectly matched layer, PML, Lippman–Schwinger equation

AMS subject classifications. 65R20, 34E10, 78M10, 78M35

DOI. 10.1137/070701388

1. Introduction.

This paper deals with the calculation of resonances of thin high contrast dielectric structures. Specifically, we are motivated by recent developments in photonic band gap (PBG) devices. PBG materials are artificially created structures having a refraction index which is spatially periodic, often on the nanoscale. As the name suggests, electromagnetic waves of frequencies in a “band gap” cannot propagate within PBG materials. These materials thus offer interesting possibilities for radical manipulation of light through introduction of defects, hence the increasing interest in them.

While the existence of band gaps has been definitively demonstrated for certain infinite periodic structures, practical PBG devices are of finite extent. When a band gap exists in a medium of infinite extent, it is possible to create a so-called defect mode, which is a standing wave of frequency in the band gap, by introducing a localized defect into the medium [6]. Such a mode corresponds to an eigenfunction of the partial differential equation governing the system. However, when the medium is of finite extent, such an eigenvalue no longer exists, but instead we may have a localized resonance mode.

[*] Received by the editors August 27, 2007; accepted for publication (in revised form) April 22, 2008; published electronically August 1, 2008. This work was supported by the National Science Foundation under grants SCREMS-0619080, DMS-0749396, DMS-0713833, and DMS-0504185.
http://www.siam.org/journals/siap/69-1/70138.html

[†] Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (jayg@math.ufl.edu).

[‡] Department of Mathematics, Drexel University, Philadelphia, PA 19104-2875 (moskow@math.drexel.edu).

[§] School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (santosa@math.umn.edu).

A particularly interesting class of PBG structures are high index thin film devices where light is confined to the film by total internal reflection, and the PBG effect is achieved by drilling an array of air holes. Examples of such thin film devices can be found in [5, 19, 20]. The present work is aimed at calculating resonances for such structures. The dielectric properties of these structures are the restrictions of a periodic function to a bounded thin region in $\mathbb{R}^3$. Additionally they have a local "defect," i.e., a break in the periodic pattern. We want to identify resonance modes that are localized near the defect region, if any.

To model such structures, we consider the simplest equation for time harmonic wave propagation, namely, the Helmholtz equation

$$\Delta u + k^2 \varepsilon(\boldsymbol{x})\, u = 0, \qquad \boldsymbol{x} \in \mathbb{R}^n, \tag{1.1}$$

where $n = 2$ or 3, and u and ε are functions of $\boldsymbol{x}$ in $\mathbb{R}^n$. By an abuse in terminology we will call $\varepsilon(\boldsymbol{x})$ the dielectric constant. The geometry is captured by the variable coefficient $\varepsilon(\boldsymbol{x})$ which is set to unity in the background (air). The function $\varepsilon(\boldsymbol{x}) - 1$ is assumed to have compact support. To find resonances, we must find a nontrivial "radiating" mode u and a complex number $\lambda \equiv k^2$ such that

$$-\Delta u = \lambda \varepsilon(\boldsymbol{x}) u \qquad \text{in } \mathbb{R}^n. \tag{1.2}$$

When k is real, a mathematically precise form of the condition that u is "radiating" (or "outgoing") is the well-known Sommerfeld radiation condition at infinity. Writing (1.2) together with the Sommerfeld condition as $Au = \lambda Bu$, the resolvent $(A - \lambda B)^{-1}$ is well defined for λ in the positive real axis, because the Sommerfeld condition gives uniqueness of Helmholtz solutions. When k (or λ) is complex, one way to make the "outgoing" condition precise is by analytic continuation from the positive real axis. For instance, for a slightly different scattering problem studied in [13], the resolvent was proved to be a meromorphic function of λ and resonances were characterized as its poles occurring in the lower half of the complex plane.

The resonance modes u satisfying equations like (1.2) are sometimes also known as *quasi-normal modes* [12]. They are nonphysical and grow exponentially when k is in the fourth quadrant. To give a physical interpretation of resonance, we must go to the time domain and consider

$$\Delta U - \varepsilon(\boldsymbol{x})\, U_{tt} = 0, \qquad \boldsymbol{x} \in \mathbb{R}^n.$$

Resonance in this context is a time-dependent solution of the equation that resembles a standing wave except for the amplitude decay. Such a solution, especially when the decay is slow, is well captured by a superposition of quasi-normal modes with e^{ikt} modulation [12]. Generally these slowly decaying resonance modes are computed by using finite difference time domain (FDTD) methods, which are computationally intensive. In this work we propose two other ways to calculate resonances.

For thin devices, we propose a direct approach based on the Lippman–Schwinger reformulation. We assume a structure which fits into the following high contrast model, namely, the dielectric occupies the region $\Omega \times (-h/2, h/2)$. In three dimensions, Ω is a bounded planar domain, while in two space dimensions, Ω is bounded domain on the real line. Thus, in either case, Ω is of one space dimension less than n. Let $\boldsymbol{x} = (x, z)$ for $x \in \mathbb{R}^2$ and $z \in \mathbb{R}$. We assume that

$$\varepsilon(x, z) = \begin{cases} \dfrac{\varepsilon_0(x)}{h} & \text{if } |z| < h/2 \text{ and } x \in \Omega, \\ 1 & \text{otherwise.} \end{cases} \tag{1.3}$$

Note that whenever we have a membrane whose dielectric properties vary negligibly across its thickness h, we can satisfy this assumption by setting ε_0 to $h\varepsilon$. In [14], we studied scattering by this type of structure and found a limiting (i.e., effective) problem as $h \to 0$, with a correction term that improved the approximation. In this paper, we study the related resonance problem.

We *define* the resonant frequency k as a number in the complex plane for which there is a nontrivial resonance mode u satisfying

$$u(x,z) = \lambda \int_\Omega \int_{-h/2}^{h/2} \left(1 - \frac{\varepsilon_0(x')}{h}\right) G_\lambda(x,z,x',z') u(x',z')\, dz' dx'. \tag{1.4}$$

Here $\lambda \equiv k^2$ (which we will call the resonance value) and G is the Helmholtz fundamental solution (in two or three dimensions). One can show by variational arguments that such a λ must necessarily be in the lower half plane, and hence k must be in the fourth quadrant. This integral equation is arrived at from (1.1) by the same standard manipulations used in deriving the Lippman–Schwinger equation for scattering problems (see Theorem 8.3 of [4]). However, since λ has negative imaginary part, the solutions to (1.4) are exponentially growing at infinity, and such manipulations are only formal. Nevertheless this suggests that the definition of resonances using (1.4) is equivalent to (1.2). As another way to see why this is the case [8], consider the operator

$$(\Delta + k^2\epsilon)$$

for real k. One can then write the outgoing Green's function to characterize the inverse of this operator with Sommerfeld radiation conditions,

$$R(k) = (\Delta + k^2\epsilon)^{-1}.$$

If one continues this operator to negative complex k, the classical definition of resonance is its poles. Now we can rewrite this operator as

$$\begin{aligned} R(k) &= (\Delta + k^2 + k^2(\epsilon - 1))^{-1} \\ &= \left[(\Delta + k^2)(I + (\Delta + k^2)^{-1}k^2(\epsilon - 1))\right]^{-1} \\ &= \left[I + (\Delta + k^2)^{-1}k^2(\epsilon - 1)\right]^{-1} (\Delta + k^2)^{-1}. \end{aligned}$$

Since the term $(\Delta + k^2)^{-1}$ is characterized by the free space Green's function, it has no poles. So, the poles of $R(k)$ are exactly where

$$I + (\Delta + k^2)^{-1}k^2(\epsilon - 1)$$

has a null space, i.e., where (1.4) has a solution. Note that our theoretical analysis neither refers to nor deals with this equivalence. Indeed, our analysis takes (1.4) as the definition of resonances and proceeds to examine how such resonance values vary with h. Since G depends on λ, (1.4) is a *nonlinear* eigenvalue problem.

Another approach for numerical approximation of resonances is to directly approximate the eigenvalue problem in (1.1) with an outgoing boundary condition. A standard technique to handle outgoing boundary conditions at infinity is by introducing a perfectly matched layer (PML) [1] away from all inhomogeneities and eventually truncating the layer to obtain a finite computational domain. This suggests the use of PML for computing resonances by solving a linear eigenvalue problem in a truncated domain. We investigate this approach numerically, comparing the results with an exact solution as well as with approximations from the asymptotic approach. For the case of thin structures we can use both approaches to validate one another. A significant finding of this paper is that with both the asymptotic method and the PML calculations we can reproduce the high quality factor (low loss) resonance mode found in [5] by FDTD methods.

The next section contains a derivation and analysis of an asymptotic approximation to resonance solutions of (1.4) with respect to the thickness parameter h. Within this section, we prove convergence of the related operators, prove convergence of the resonance values, and then finally derive a correction term for the resonances utilizing an eigenvalue approximation theorem of Osborn [16]. Section 3 contains a numerical study of both asymptotic/nonlinear eigenvalue and PML approaches to find the resonance solutions. In section 3.1, we find exact solutions for the resonances of a disk and use them to analyze the convergence of PML solutions. In section 3.2 we return to a thin, high contrast structure. We compute resonances with both PML and asymptotics for the same problem and compare the results. In section 3.3 we study a thin periodic structure with a defect from [5], which was previously found to exhibit a localized low loss mode. The concluding section summarizes our results.

2. An asymptotic limit. In this section we develop an asymptotic approach to the resonance approximation for these thin, high contrast structures. The resonance problem is then formulated in terms of operator equations, and we prove operator convergence, that is, we show that the operators depend continuously on the thickness and frequency parameters. In the subsections that follow, we show that the resonance values converge and prove an error estimate. In the case of a simple resonance value, we introduce a correction term that increases the accuracy of the asymptotic approximation.

Assume that we have a dielectric with geometry defined by (1.3). A resonance value λ_h is a complex number for which there is a nontrivial function u_h satisfying

$$u_h(x,z) = \lambda_h \int_\Omega \int_{-h/2}^{h/2} \left(1 - \frac{\varepsilon_0(x')}{h}\right) G_{\lambda_h}(x,z,x',z') u_h(x',z') dz' dx', \tag{2.1}$$

where ε_0 is assumed to be piecewise continuous and G is the Helmholtz fundamental solution (in two or three dimensions) with complex $\lambda_h = k^2$. That is, when $n = 3$,

$$G_\lambda(x,z,x',z') = -\frac{1}{4\pi} \frac{e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+|z-z'|^2}}}{\sqrt{|x-x'|^2+|z-z'|^2}},$$

and when $n = 2$,

$$G_\lambda(x,z,x',z') = -\frac{i}{4} H_0^{(1)}(\sqrt{\lambda}\sqrt{|x-x'|^2+|z-z'|^2}),$$

where $H_0^{(1)}$ is a Hankel function of the first kind. We note that we are taking the branch of the square root in the complex plane for which the cut is on the negative real axis, and hence there is analyticity away from this cut. With the scaling in the z direction, $z = h\zeta$, let

$$\tilde{u}_h(x,\zeta) = u_h(x,z)$$

to obtain

$$\tilde{u}_h(x,\zeta) = \lambda_h \int_\Omega \int_{-1/2}^{1/2} (h - \varepsilon_0(x')) \, G_{\lambda_h}(x,h\zeta,x',h\zeta') \tilde{u}_h(x',\zeta') d\zeta' dx'. \tag{2.2}$$

Now, if we let $h \to 0$, this leads us to guess the limiting resonance problem: Find nontrivial solutions (u_0, λ_0) to

$$u_0(x) = -\lambda_0 \int_\Omega \varepsilon_0(x') G_{\lambda_0}(x, 0, x', 0) u_0(x') dx'. \tag{2.3}$$

Although still a nonlinear eigenvalue problem, this has one dimension less than we started with.

In order to analyze the validity of this asymptotic limit, it is useful to express these problems in operator form. Let S be the scaled, fixed domain

$$S = \Omega \times [-1/2, 1/2].$$

Consider, for $v \in L^2(S)$, the operators $T_h(\lambda)$ and $T_0(\lambda)$, with complex parameter λ, are defined by

$$T_h(\lambda)v = \int_{-1/2}^{1/2} \int_\Omega (h - \varepsilon_0(x')) G_\lambda(x, h\zeta, x', h\zeta') v(x', \zeta') dx' d\zeta'$$

and

$$T_0(\lambda)v = -\int_{-1/2}^{1/2} \int_\Omega \varepsilon_0(x') G_\lambda(x, 0, x', 0) v(x', \zeta') dx' d\zeta'.$$

The operators $T_h(\lambda)$ and $T_0(\lambda)$ are both compact from $L^2(S)$ to $L^2(S)$ by the proof of [14, Lemma 2]. (Unlike in that lemma, since here λ is not necessarily real, we are not ensured the invertibility of $(I - \lambda T_h(\lambda))$ or $(I - \lambda T_0(\lambda))$, hence the presence of resonance values.) We say that λ_h is a resonance value of T_h if there exists nontrivial $u_h \in L^2(S)$ such that

$$u_h = \lambda_h T_h(\lambda_h) u_h.$$

Similarly, λ_0 is a resonance value of T_0 if there exists nontrivial u_0 such that

$$u_0 = \lambda_0 T_0(\lambda_0) u_0.$$

The operators T_h, T_0 are compact on $C^0(S)$ as well as $L^2(S)$, but here we will use $L^2(S)$ for its Hilbert space structure. We use $\langle , \rangle$ to denote the standard $L^2(S)$ inner product over $\mathbb{C}$:

$$\langle u, v \rangle := \int_S u\overline{v},$$

where $\overline{v}$ is the complex conjugate of v.

2.1. Operator convergence. We first prove a lemma showing convergence of the fundamental solutions when $n = 3$. The same result also holds for $n = 2$. This is an extension of [14, Lemma 1] to complex λ. Here we also give the explicit dependence of the constant on λ. Recall the definition of the scaled domain

$$S = \Omega \times (-1/2, 1/2).$$

LEMMA 2.1. *There exists a constant C independent of h, ζ', and λ, such that*

$$\sup_{(x,\zeta)\in S} \int_\Omega |G_\lambda(x, 0, x', 0) - G_\lambda(x, h\zeta, x', h\zeta')| dx' \le Ch(1 + |\sqrt{\lambda}|) e^{|\mathcal{I}m\sqrt{\lambda}| diam(\Omega_h)}.$$

Proof. The difference of these fundamental solutions can be written as

$$\begin{aligned}
&G(x, h\zeta, x', h\zeta') - G(x, 0, x', 0)\\
&= \frac{1}{4\pi}\frac{e^{i\sqrt{\lambda}|x-x'|}}{|x-x'|} - \frac{1}{4\pi}\frac{e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\\
&= \frac{1}{4\pi}e^{i\sqrt{\lambda}|x-x'|}\left[\frac{1}{|x-x'|} - \frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right]\\
&\quad + \frac{1}{4\pi}\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\left[e^{i\sqrt{\lambda}|x-x'|} - e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right]. \qquad (2.4)
\end{aligned}$$

We first work on the second term on the right-hand side of (2.4). By a standard Taylor expansion,

$$e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}} = e^{i\sqrt{\lambda}|x-x'|} + i\sqrt{\lambda}\left(\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2} - |x-x'|\right)e^{i\sqrt{\lambda}\xi}$$

for some ξ between $|x-x'|$ and $\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}$. Since we know that for $(x,\zeta) \in S$,

$$\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2} - |x-x'| \le h,$$

we obtain

$$\begin{aligned}
\left|e^{i\sqrt{\lambda}|x-x'|} - e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right| &\le |\sqrt{\lambda}|he^{|\mathcal{I}m\sqrt{\lambda}|\xi}\\
&\le |\sqrt{\lambda}|he^{|\mathcal{I}m\sqrt{\lambda}|\operatorname{diam}(\Omega_h)}. \qquad (2.5)
\end{aligned}$$

Also,

$$\frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}} \le \frac{1}{|x-x'|},$$

which is integrable with respect to x' on Ω, and we have that

$$\int_\Omega \frac{dx'}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}$$

is bounded independently of h, ζ', λ, and $(x,z) \in S$. This along with (2.5) gives that we can choose C independent of h, ζ' and $(x,\zeta) \in S$ such that

$$\int_\Omega \frac{1}{4\pi}\frac{|e^{i\sqrt{\lambda}|x-x'|} - e^{i\sqrt{\lambda}\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}|}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}dx' \le Ch|\sqrt{\lambda}|e^{|\mathcal{I}m\sqrt{\lambda}|\operatorname{diam}(\Omega_h)}. \qquad (2.6)$$

The integral of the first term on the right-hand side of (2.4) can be bounded,

$$\begin{aligned}
&\int_\Omega \left|\frac{1}{4\pi}e^{i\sqrt{\lambda}|x-x'|}\left[\frac{1}{|x-x'|} - \frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right]\right|dx'\\
&\le \frac{1}{4\pi}e^{|\mathcal{I}m\sqrt{\lambda}|\operatorname{diam}(\Omega_h)}\int_\Omega\left|\frac{1}{|x-x'|} - \frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right|dx'\\
&= \frac{1}{4\pi}e^{|\mathcal{I}m\sqrt{\lambda}|\operatorname{diam}(\Omega_h)}\int_\Omega\left(\frac{1}{|x-x'|} - \frac{1}{\sqrt{|x-x'|^2+h^2|\zeta-\zeta'|^2}}\right)dx',
\end{aligned}$$

since the integrand is nonnegative. Now choose R large enough so that if $B_R(x)$ is the ball of radius R centered at x in $\mathbb{R}^2$,

$$\Omega \subset B_R(x)$$

for all $x \in \Omega$. Then the integral over Ω above is bounded by

$$\leq \int_{B_R(x)} \left(\frac{1}{|x-x'|} - \frac{1}{\sqrt{|x-x'|^2 + h^2|\zeta - \zeta'|^2}} \right) dx'.$$

Change to polar coordinates centered at x with

$$r = |x - x'|.$$

The integral transforms to

$$\begin{aligned} &= 2\pi \int_0^R \left(\frac{1}{r} - \frac{1}{\sqrt{r^2 + h^2|\zeta-\zeta'|^2}} \right) \\ &= 2\pi \left[R - \sqrt{R^2 + h^2|\zeta - \zeta'|^2} + h|\zeta - \zeta'| \right] \end{aligned}$$

by direct calculation. One can see clearly that this quantity is then $O(h)$, where the constant is independent of $(x, \zeta) \in S$, λ, and $\zeta' \in (-1/2, 1/2)$. This, combined with (2.4) and the estimate (2.6), proves the lemma. □

Next we show that the operators depend continuously on the parameters h and λ.

PROPOSITION 2.1. *Assume we have a sequence of pairs* $\{h_j, \lambda_j\}$, *where* $h_j \in \mathbb{R}$, *the* λ_j *are in the complex plane with the negative real axis and the origin removed, i.e.,* $\lambda_j \in \mathbb{C} \setminus \{\mathbb{R}^- \cup \{0\}\}$, *and for which* $\lambda_j \to \lambda_0$ *for some* $\lambda_0 \in \mathbb{C} \setminus \{\mathbb{R}^- \cup \{0\}\}$, *and* $h_j \to 0$ *as* $j \to \infty$. *Then*

$$T_{h_j}(\lambda_j) \to T_0(\lambda_0)$$

in the operator norm on $L^2(S)$ *as* $j \to \infty$. *Furthermore, for* j *large enough there exists* C *independent of* j *such that*

$$\|T_{h_j}(\lambda_j) - T_0(\lambda_0)\| \leq C\left(h_j + |\lambda_j - \lambda_0|\right).$$

Proof. Consider, for $v \in L^2(S)$,

$$\left(T_{h_j}(\lambda_j) - T_0(\lambda_0)\right) v = \left(T_{h_j}(\lambda_j) - T_0(\lambda_j)\right) v + \left(T_0(\lambda_j) - T_0(\lambda_0)\right) v.$$

We will expand out the second term on the right-hand side:

$$\left(T_0(\lambda_j) - T_0(\lambda_0)\right) v = \int_{-1/2}^{1/2} \int_\Omega \varepsilon_0(x')(G_{\lambda_0}(x,0,x',0) - G_{\lambda_j}(x,0,x',0)) v(x',\zeta') dx' d\zeta'.$$

We can use the mean value theorem; for $|x - x'| \neq 0$,

$$\begin{aligned} G_{\lambda_j}(x,0,x',0) - G_{\lambda_0}(x,0,x',0) &= \frac{1}{4\pi} \frac{e^{i\sqrt{\lambda_0}|x-x'|}}{|x-x'|} - \frac{1}{4\pi} \frac{e^{i\sqrt{\lambda_j}|x-x'|}}{|x-x'|} \\ &= (\lambda_0 - \lambda_j) \frac{i}{8\pi\sqrt{\eta}} e^{i\sqrt{\eta}|x-x'|} \end{aligned}$$

for some η on the line in $\mathbb{C}$ joining λ_0 and λ_j. So, since λ_0 is bounded away from the negative real axis, for large enough j we have

$$\|G_{\lambda_0}(x,0,x',0) - G_{\lambda_j}(x,0,x',0)\|_\infty \le C|\lambda_0 - \lambda_j|$$

for some C independent of j. From this we easily obtain

$$\|\,(T_0(\lambda_j) - T_0(\lambda_0))\,v\|_{L^2(S)} \le C|\lambda_0 - \lambda_j|\|v\|_{L^2(S)}. \tag{2.7}$$

Now, for the other term,

$$\begin{aligned}(T_{h_j}(\lambda_j) - T_0(\lambda_j))v = h_j \int_S G_{\lambda_j}(x, h_j\zeta, x', h_j\zeta')v(x',\zeta')dx'd\zeta' \\ - \int_S \varepsilon_0(x')\left[G_{\lambda_j}(x,h_j\zeta,x',h_j\zeta') - G_{\lambda_j}(x,0,x',0)\right] v(x'\zeta')dx'd\zeta'.\end{aligned} \tag{2.8}$$

Now, since G_{λ_j} is a kernel which is bounded in L^1 independently of h, it follows from the generalized Young inequality [7] that the function

$$w(x,\zeta) = \int_S G_{\lambda_j}(x,h_j\zeta,x',h_j\zeta')v(x',\zeta')dx'd\zeta'$$

satisfies

$$\|w\|_{L^2(S)} \le C\|v\|_{L^2(S)},$$

which shows that the first term in (2.8) is $O(h_j)$. For the second term, we appeal to Lemma 2.1, which tells us that the kernel difference can be bounded:

$$\|G_{\lambda_j}(x,h_j\zeta,x',h_j\zeta') - G_{\lambda_j}(x,0,x',0)\|_{L^1(S)} \le Ch_j,$$

where C is independent of j. Since ε_0 is bounded in L^∞, again using the generalized Young and triangle inequalities in (2.8), we obtain

$$\|(T_{h_j}(\lambda_j) - T_0(\lambda_j))v\|_{L^2(S)} \le Ch_j\|v\|_{L^2(S)}.$$

Combining this with (2.7), the result follows. □

2.2. Convergence of the resonance values. Define the modified resolvent type operator-valued functions on $\mathbb{C}$

$$R_h(\lambda) = (I - \lambda T_h(\lambda))^{-1}$$

and

$$R_0(\lambda) = (I - \lambda T_0(\lambda))^{-1}.$$

An important note is that if $R_h(\lambda)$ does not exist as a bounded linear operator from $L^2(S)$ to itself, then λ is a resonance value of T_h. This is because if $R_h(\lambda)$ does not exist, then $1/\lambda$ is in the spectrum of the compact operator $T_h(\lambda)$. Hence $1/\lambda$ must be an eigenvalue, and $(I - \lambda T_h(\lambda))$ must have a nontrivial and finite-dimensional null space. The same holds for the limiting operator $R_0(\lambda)$.

In the following theorem, we show that, with an assumption of nonzero residue, the resonance values converge.

THEOREM 2.2. *Assume that λ_0 is a resonance value of T_0, and that R_0 and R_h are meromorphic in some region of $\mathbb{C}$ containing λ_0. Assume also that R_0 has nonzero residue at λ_0. Then for any ball B around λ_0, there exists $h_0 > 0$ such that T_h has a resonance in B for all $h < h_0$. Conversely, if $\{\lambda_h\}$ is a sequence of resonance values of T_h that converges as $h \to 0$, the limit is a resonance value of T_0.*

Proof. We first note that we know from [14] that $\lambda_0 \notin \mathbb{R}$. So, we can choose B, a ball around λ_0 which does not intersect the negative real axis $\mathbb{R}^-$, and such that T_0 has no other resonance values in $\overline{B}$.

We will also use a well-known result about the inverses of perturbed operators (see, for example, [10, p. 31]): If $S - T = A$ and T^{-1} exists, then for $\|A\| < \frac{1}{\|T^{-1}\|}$, S^{-1} exists and

$$\|S^{-1} - T^{-1}\| \leq \frac{\|A\|\|T^{-1}\|^2}{1 - \|A\|\|T^{-1}\|}. \tag{2.9}$$

Apply this, with

$$S = I - \lambda T_h(\lambda),$$

$$T = I - \lambda T_0(\lambda),$$

to get

$$\|R_h(\lambda) - R_0(\lambda)\| \leq \frac{\|\lambda(T_0(\lambda) - T_h(\lambda))\|\|R_0(\lambda)\|^2}{1 - \|\lambda(T_0(\lambda) - T_h(\lambda))\|\|R_0(\lambda)\|},$$

which, from Proposition 2.1, yields

$$\|R_h(\lambda) - R_0(\lambda)\| \leq \frac{Ch\|R_0(\lambda)\|^2}{1 - Ch\|R_0(\lambda)\|} \tag{2.10}$$

for C independent of h, for h small enough. The constant C does in general depend on λ, but for λ on a compact subset of $\mathbb{C}$ bounded away from the real line, C can be chosen independent of λ.

Let $\Gamma = \partial B$, positively oriented. By the choice of B, Γ does not intersect with $\mathbb{R}^-$, and λ_0 is the only pole of R_0 in the closed disk. Then $R_0(\lambda)$ is continuous with respect to λ on Γ, and hence $\|R_0(\lambda)\|$ is uniformly bounded for λ on Γ. Using (2.10), we have that

$$R_h(\lambda) \to R_0(\lambda)$$

in norm as $h \to 0$, uniformly for $\lambda \in \Gamma$. This implies that the operator-valued integral

$$\frac{1}{2\pi i}\int_\Gamma R_h(\lambda) d\lambda \to \frac{1}{2\pi i}\int_\Gamma R_0(\lambda) d\lambda$$

in norm as $h \to 0$. From the residue theorem, the integral

$$\frac{1}{2\pi i}\int_\Gamma R_0(\lambda) d\lambda$$

gives us the coefficient of the $(\lambda - \lambda_0)^{-1}$ term in the Laurent series expansion for $R_0(\lambda)$, which by assumption is nonzero. Hence the integrals must all be nonzero for

h small enough. This means that all R_h must have at least one pole in B for h small enough. That is, for h small enough, all T_h have a resonance value in B. This proves the first part of the statement of the proposition. For the converse, if λ_0 is not a resonance value of T_0, then $R_0(\lambda)$ exists in some neighborhood of λ_0. The formula (2.10) implies that $R_h(\lambda)$ also exists in that neighborhood for h small enough. Hence the resonance values of T_h are bounded away from λ_0 for h small enough. □

Some remarks about the assumptions in this theorem:

- The operator functions $I - \lambda T_h(\lambda)$ and $I - \lambda T_0(\lambda)$ are analytic with respect to λ away from the negative real axis. This, combined with the fact that the T's are compact, means that the inverses are meromorphic.
- If λ_0 is a resonance of T_0, then the classical resolvent of $T_0(\lambda_0)$, given by $(zI - T_0(\lambda_0))^{-1}$, automatically has nonzero residue at $z = \frac{1}{\lambda_0}$; its residue is the projection onto the generalized eigenspace [10]. It is not clear how the residue of R_0 relates to the nonlinear eigenspace. However, if λ_0 is a simple pole, this is a special case of nonzero residue.
- We do not need the assumption about nonzero residue for the converse.

2.3. A higher order correction. Once we know that we have a convergent sequence of resonance values as $h \to 0$, we can use standard eigenvalue perturbation theorems. In the resonance value expansion, we employ a result of Osborn [16] which is valid for nonself-adjoint operators and also yields a correction term. The actual result in [16] is more general, but we state it here for the case of norm convergence on a Hilbert space.

Suppose X is a Hilbert space and $T_n : X \to X$ is a sequence of compact linear operators such that $T_n \to T$ in norm. It then follows that the adjoint operators also converge in norm. Let μ be a nonzero eigenvalue of T of algebraic multiplicity m. It is well known that for n large enough there exist m eigenvalues of T_n, $\mu_1^n, \ldots, \mu_m^n$ (counted according to algebraic multiplicity) such that $\mu_j^n \to \mu$ as $n \to \infty$ for each $1 \le j \le m$.

Let E be the spectral projection onto the generalized eigenspace of T corresponding to eigenvalue μ. The space X can be decomposed in terms of the range and null space of E: $X = R(E) \oplus N(E)$.

THEOREM 2.3 (Osborn). *Let $\phi_1, \phi_2, \ldots, \phi_m$ be a normalized basis for $R(E)$. Then there exists a constant C such that*

$$\left| \mu - \frac{1}{m}\sum_{j=1}^{m} \mu_j^n - \frac{1}{m}\sum_{j=1}^{m} \langle (T - T_n)\phi_j, \phi_j\rangle \right| \le C \|(T - T_n)|_{R(E)}\| \cdot \|(T^* - T_n^*)|_{R(E^*)}\|.$$

To simplify the statement of the following theorem, we define the lower-dimensional operator $DT_0(\lambda)$ by

$$DT_0(\lambda)v = -\int_\Omega \varepsilon_0(x') \frac{\partial G}{\partial \lambda}(x, 0, x', 0) v(x') dx', \tag{2.11}$$

and we leave off the subscripts j for the sequence $\{h_j\}$ of values of h going to zero.

THEOREM 2.4. *Assume we have a sequence $\{\lambda_h\} \in \mathbb{C}$ of resonance values of T_h for which $\lambda_h \to \lambda_0$ as $h \to 0$, where $\lambda_0 \in \mathbb{C}$ is a simple resonance value of T_0 with normalized resonance function u_0 satisfying $\lambda_0 T_0(\lambda_0)u_0 = u_0$. Assume also that*

$$\lambda_0^2 \langle DT_0(\lambda_0)u_0, u_0\rangle \ne -1. \tag{2.12}$$

Then there exists C independent of h such that

$$|\lambda_h - \lambda_0| \leq Ch,$$

and furthermore

$$\lambda_h = \lambda_0 + \lambda_0^2 \frac{\langle (T_0(\lambda_0) - T_h(\lambda_0))u_0, u_0\rangle}{1 + \lambda_0^2 \langle DT_0(\lambda_0)u_0, u_0\rangle} + O(h^2). \tag{2.13}$$

Proof. Note that

$$\lambda_h T_h(\lambda_h) u_h = u_h$$

and

$$\lambda_0 T_0(\lambda_0) u_0 = u_0;$$

that is, $\frac{1}{\lambda_h}$ is an eigenvalue of $T_h(\lambda_h)$ and $\frac{1}{\lambda_0}$ is an eigenvalue of $T_0(\lambda_0)$. Also, from Proposition 2.1, we know that

$$T_h(\lambda_h) \to T_0(\lambda_0)$$

in the operator norm. So, what we have are the eigenvalues of a convergent sequence of compact operators. These operators, $\{T_h(\lambda_h), T_0(\lambda_0)\}$, are not self-adjoint, but it follows from the norm convergence that the adjoints also converge:

$$T_h^*(\lambda_h) \to T_0^*(\lambda_0)$$

in the operator norm, with the same norm error. Since we assume that $\frac{1}{\lambda_0}$ is a simple eigenvalue of $T_0(\lambda_0)$, Theorem 2.3 yields

$$\begin{aligned} &\left| \frac{1}{\lambda_0} - \frac{1}{\lambda_h} - \langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle \right| \\ &\qquad \leq \|(T_0(\lambda_0) - T_h(\lambda_h))u_0\| \cdot \|(T_0^*(\lambda_0) - T_h^*(\lambda_h))u_0\|. \end{aligned} \tag{2.14}$$

Since

$$\|T_0^*(\lambda_0) - T_h^*(\lambda_h)\| = \|T_0(\lambda_0) - T_h(\lambda_h)\|,$$

we have from Proposition 2.1

$$\left| \frac{1}{\lambda_0} - \frac{1}{\lambda_h} - \langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle \right| \leq C\left(h + |\lambda_0 - \lambda_h|\right)^2.$$

If we multiply everything by $\lambda_0\lambda_h$,

$$|\lambda_h - \lambda_0 - \lambda_0\lambda_h \langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle| \leq C\left(h + |\lambda_0 - \lambda_h|\right)^2,$$

which we manipulate to get

$$\begin{aligned} \lambda_h = \lambda_0 &+ \lambda_0^2 \langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle \\ &+ \lambda_0(\lambda_h - \lambda_0)\langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle + O\left((h + |\lambda_h - \lambda_0|)^2\right). \end{aligned}$$

Again using Proposition 2.1,

$$\lambda_h = \lambda_0 + \lambda_0^2 \langle (T_0(\lambda_0) - T_h(\lambda_h))u_0, u_0\rangle + O\left((h + |\lambda_h - \lambda_0|)^2\right). \tag{2.15}$$

Now, since the correction term above depends on λ_h, we need to expand the term further. We can write

$$T_0(\lambda_0) - T_h(\lambda_h) = (T_0(\lambda_0) - T_h(\lambda_0)) + (T_h(\lambda_0) - T_h(\lambda_h)) \tag{2.16}$$

and compute

$$(T_h(\lambda_0) - T_h(\lambda_h)) = \int_S (h - \varepsilon_0(x'))(G_{\lambda_0} - G_{\lambda_h})(x, h\zeta, x', h\zeta')u_0(x')d\zeta' dx'.$$

Note that since the range of T_0 contains only functions that are independent of ζ, u_0 must be independent of ζ. Since we are bounded away from the negative real axis, G is analytic with respect to λ, and so by standard Taylor expansion we obtain

$$\begin{aligned}(T_h(\lambda_0) - T_h(\lambda_h)) = &\int_S (h - \varepsilon_0(x'))(\lambda_0 - \lambda_h)\frac{\partial G}{\partial \lambda}\Big|_{\lambda=\lambda_0}(x, h\zeta, x', h\zeta')u_0(x')dx'd\zeta' \\ &+ O(|\lambda_0 - \lambda_h|^2).\end{aligned}$$

Note that the integrand is now continuous. This yields, after expanding the exponential or Hankel function about $h = 0$,

$$\begin{aligned}(T_h(\lambda_0) - T_h(\lambda_h)) = &\int_S \varepsilon_0(x')(\lambda_h - \lambda_0)\frac{\partial G}{\partial \lambda}\Big|_{\lambda=\lambda_0}(x, 0, x', 0)u_0(x')dx'd\zeta' \\ &+ O\left((h + |\lambda_0 - \lambda_h|)^2\right).\end{aligned} \tag{2.17}$$

Note the above integrand is independent of ζ'. Combining (2.17), (2.16), and (2.15),

$$\lambda_h = \lambda_0 + \lambda_0^2\langle (T_0(\lambda_0) - T_h(\lambda_0))u_0, u_0\rangle - \lambda_0^2(\lambda_h - \lambda_0)\langle DT_0(\lambda_0)u_0, u_0\rangle + O\left((h + |\lambda_0 - \lambda_h|)^2\right),$$

where DT_0 is defined by (2.11). We now collect terms for $(\lambda_h - \lambda_0)$ to get

$$(\lambda_h - \lambda_0)\left(1 + \lambda_0^2\langle DT_0(\lambda_0)u_0, u_0\rangle\right) = \lambda_0^2\langle (T_0(\lambda_0) - T_h(\lambda_0))u_0, u_0\rangle + O\left((h + |\lambda_0 - \lambda_h|)^2\right).$$

At this point we need to use the assumption (2.12) to obtain

$$\lambda_h = \lambda_0 + \frac{\lambda_0^2\langle (T_0(\lambda_0) - T_h(\lambda_0))u_0, u_0\rangle}{1 + \lambda_0^2\langle DT_0(\lambda_0)u_0, u_0\rangle} + O\left((h + |\lambda_0 - \lambda_h|)^2\right).$$

Recall that by the proof of Proposition 2.1,

$$\|T_0(\lambda_0) - T_h(\lambda_0)\| = O(h)$$

in the operator norm, and so we have

$$\lambda_h - \lambda_0 = O(h) + O\left((h + |\lambda_0 - \lambda_h|)^2\right).$$

Since we assume that $\lambda_h - \lambda_0 \to 0$, this can only hold if

$$\lambda_h - \lambda_0 = O(h).$$

This completes the proof. □

Remark 2.1. It is possible that the hypothesis (2.12) is related to the residue of the generalized resolvent $R_0(\lambda)$. In particular, we conjecture that for a simple resonance value, (2.12) holds if and only if the residue is nonzero, i.e., the requirement for convergence in Theorem 2.2.

The correction term above is not difficult to compute since it involves only applying integral operators to the limiting resonance function u_0. However, one may want to have an expression of the form

$$\lambda_h \approx \lambda_0 + h\lambda^{(1)}$$

in which $\lambda^{(1)}$ is independent of h. We had studied the numerator of the correction exactly in [14] for the case when λ_0 was real. By that same analysis,

$$\begin{aligned}(T_0(\lambda_0) - T_h(\lambda_0))u_0 &= -h\int_\Omega\int_{-1/2}^{1/2} G_{\lambda_0}(x, h\zeta, x', h\zeta')u_0 d\zeta' dx' \\ &+ \int_\Omega\int_{-1/2}^{1/2} \varepsilon_0(x')(G_{\lambda_0}(x, h\zeta, x', h\zeta') - G_{\lambda_0}(x, 0, x', 0))u_0(x')d\zeta' dx' \\ &= -h\int_\Omega G_{\lambda_0}(x, 0, x', 0)u_0(x')dx' \\ &\qquad + h\frac{\varepsilon_0(x)u_0(x)}{2}\left(\zeta^2 + \frac{1}{4}\right) + o(h).\end{aligned} \tag{2.18}$$

This yields the following corollary.

Corollary 2.5. *Assume the hypotheses as in Theorem* 2.4. *Then*

$$\lambda_h = \lambda_0 + h\lambda_0^2\frac{\langle g, u_0\rangle}{1 + \lambda_0^2\langle DT_0(\lambda_0)u_0, u_0\rangle} + o(h),$$

where

$$g(x) = -\int_\Omega G_{\lambda_0}(x, 0, x', 0)u_0(x')dx' + \frac{\varepsilon_0(x)u_0(x)}{2}\left(\zeta^2 + \frac{1}{4}\right).$$

Note, however, that here the error is no longer guaranteed to be $O(h^2)$.

3. Numerical techniques. In this section, we investigate two numerical approaches for the computation of resonances of thin membranes. The first is via Berenger's perfectly matched layer (PML) [1]. The second is a collocation discretization of the integral equation formulation combined with the asymptotics developed in the previous section. The numerical analyses of both of these approaches for resonance computation are presently open. Nonetheless, considerable insight into these computational approaches can be gained by comparing them with each other.

First, we will exhibit an example with a disk where we can compute resonances exactly. We will compute the approximations to these exact resonances using the PML approach and compare. Since no error analysis is known for the PML eigenvalue approximations, this will serve as validation of our first approach. Note that although PML has been increasingly used for computation of open resonances [9, 11, 17], we have not been able to locate a comparison of approximate and exact resonances in the literature—another reason for including such a comparison here.

Next, we will examine thin, high contrast homogeneous structures. Here will investigate the asymptotic integral equation approach along with PML. While the PML approach reduces to a *large sparse generalized eigenvalue computation*, the second approach yields a *small dense nonlinear eigenvalue problem.* We will establish that the asymptotics are sound by testing the results against the ones obtained from PML.

Finally, we will use both the PML approach and asymptotics to compute a resonance mode found in [5] for a periodic structure with a defect. The mode is localized near the defect; that is, it exhibits photonic band gap–type behavior. Also, it has a high quality factor, indicating that in the time domain its decay is slow. In [5] the mode was computed using an FDTD (finite difference time domain) method, as is typically the case.

3.1. A disk: Exact resonances and PML validation. We will now calculate the first few resonance modes of a circular homogeneous dielectric disk of radius a having (constant) permittivity ε_d placed in an infinite vacuum. If the mode is written in the form

$$U(\boldsymbol{x},t) = e^{-ikt}u(\boldsymbol{x}) = \begin{cases} e^{-ikt}u^+(\boldsymbol{x}), & |\boldsymbol{x}| > a, \\ e^{-ikt}u^-(\boldsymbol{x}), & |\boldsymbol{x}| \le a, \end{cases}$$

the governing equations are

$$\Delta u^+ + k^2 u^+ = 0, \qquad \text{when } r > a \text{ (in a vacuum)}, \tag{3.1}$$

$$\Delta u^- + k^2 \varepsilon_d u^- = 0, \qquad \text{when } r \le a \text{ (in the dielectric)}, \tag{3.2}$$

$$(u^+ - u^-)\big|_{r=a} = \frac{\partial}{\partial r}(u^+ - u^-)\bigg|_{r=a} = 0 \qquad \text{(compatibility conditions)}. \tag{3.3}$$

In addition, u^+ must be an outgoing wave at infinity. We use separation of variables. Substituting $u = R(r)\Theta(\theta)$ above and proceeding in the standard way, we conclude that

$$u^+ = H^{(1)}_{\tilde{n}}(kr)(\tilde{A}e^{i\tilde{n}\theta} + \tilde{B}e^{-i\tilde{n}\theta}) \qquad \text{in a vacuum}, \tag{3.4}$$

$$u^- = J_n(k\sqrt{\varepsilon_d}r)(Ae^{in\theta} + Be^{-in\theta})n \qquad \text{in the dielectric} \tag{3.5}$$

for some integers $n, \tilde{n} = 0, 1, 2, \ldots$. Here we have picked solutions that are outgoing in the vacuum region and bounded inside the dielectric.

Now, the first transmission condition of (3.3) implies $n = \tilde{n}$,

$$\tilde{A} = A\frac{J_n(k\sqrt{\varepsilon_d}a)}{H^{(1)}_n(ka)}, \qquad \text{and} \qquad \tilde{B} = B\frac{J_n(k\sqrt{\varepsilon_d}a)}{H^{(1)}_n(ka)}, \tag{3.6}$$

and the second condition of (3.3) further yields

$$(R^-)'(a)\Theta^-(\theta) = (R^+)'(a)\Theta^+(\theta),$$

where the $-$ and $+$ signify the interior and exterior of the disk, respectively. This implies that for each n, the values of k must satisfy

$$\sqrt{\varepsilon_d}\, J_n'(k\sqrt{\varepsilon_d}a)\, H^{(1)}_n(ka) = \ (H^{(1)}_n)'(ka)\ J_n(k\sqrt{\varepsilon_d}a). \tag{3.7}$$

We have not been able to analytically solve this equation for k. However, we can obtain numerical approximations to high precision by finding the roots of the function

$$f_n = k\bigg(\sqrt{\varepsilon_d}\, J_n'(k\sqrt{\varepsilon_d}a)\, H^{(1)}_n(ka) - (H^{(1)}_n)'(ka)\ J_n(k\sqrt{\varepsilon_d}a)\bigg), \tag{3.8}$$

where we have multiplied by k to remove a singularity.

TABLE 1
A few exact resonance values $k_{n,m}$.

$k_{n,m}$	$n = 0$	$n = 1$	$n = 2$	$n = 3$
$m = 1$	$0.436676 - 0.303945i$	$1.115540 - 0.239628i$	$1.756263 - 0.174352i$	$2.384047 - 0.121696i$
$m = 2$	$1.977701 - 0.279097i$	$2.716779 - 0.266504i$	$3.404368 - 0.245056i$	$4.064044 - 0.220085i$
$m = 3$	$3.542742 - 0.276273i$	$4.298557 - 0.271174i$	$5.013898 - 0.260996i$	$5.702569 - 0.248231i$

We enumerate the exact resonance values of this problem as $k_{n,m}$, as for each n, we have a sequence of roots for (3.8), indexed by m. A few exact resonance values obtained for the case

$$a = 1 \quad \text{and} \quad \varepsilon_d = 4$$

are displayed in Table 1. Note that for $n > 0$, each resonance value $k_{n,m}$ is of multiplicity two (both A and B in (3.5) are degrees of freedom), while for $n = 0$ the resonance values $k_{0,m}$ are simple.

Now we report on some discrete approximations to these exact resonances for the disk. These approximations are computed using finite elements and PMLs. The exact problem can be cast as the eigenvalue problem of finding complex numbers $\lambda \equiv k^2$ and nontrivial eigenfunctions u satisfying

$$-\Delta u = \lambda\, \varepsilon(\boldsymbol{x}) u \qquad \text{on } \mathbb{R}^2, \qquad \text{where } \varepsilon(\boldsymbol{x}) = \begin{cases} 4 \text{ if } |\boldsymbol{x}| \le 1, \\ 1 \text{ if } |\boldsymbol{x}| > 1, \end{cases}$$

with the additional condition that u is an outgoing wave at infinity. (Note that the dielectric parameters are the same as that used to obtain Table 1.) We will use PML as an absorbing layer to exponentially damp the solution outside a fixed radius r_1, and then truncate the computational domain for some $r_3 r_1$. In the truncated finite domain, we use the finite elements as the discretization method. This is a well-known technique used for source problems [1, 2, 3] with outgoing solutions, although its applicability to eigenvalue problems is less studied.

We first briefly describe the truncated PML and its finite element approximation. Our PML parameters are closer to [2, 3] than the original ones of Berenger [1]. In the region $r < 1$, we set the actual coefficients given by our ε. The artificial coefficients forming PML are set outside radius $r = r_1 \ge 1$. In the region $r_1 < r < r_2$ we have a transitional variable coefficient, and in the region $r_2 < r < r_3$ we have a constant artificial coefficient. Define

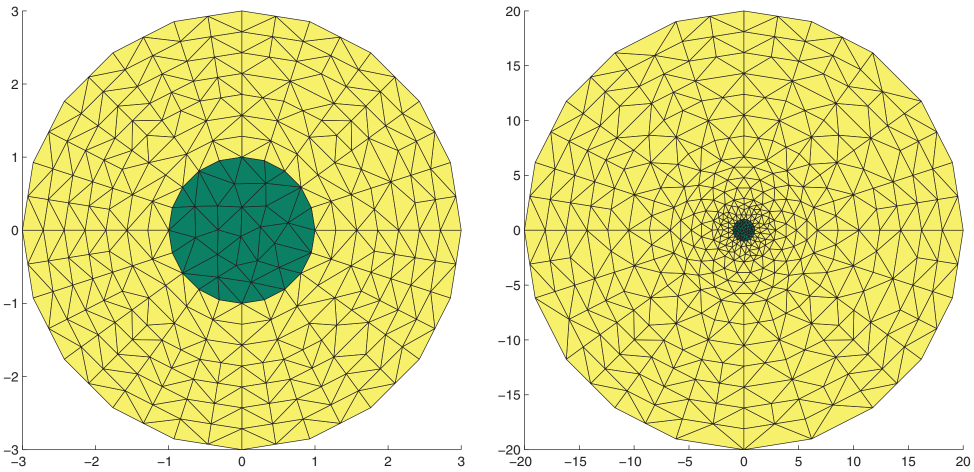
$$\text{(3.9)} \quad \tilde\sigma(r) = \begin{cases} 0 & \text{if } r < r_1, \\ \dfrac{s(r)}{s(r_2)} & \text{if } r_1 < r < r_2, \\ 1 & \text{if } r > r_2, \end{cases} \qquad \sigma(r) = \begin{cases} 0 & \text{if } r < r_1, \\ \dfrac{d}{dr}(r\tilde\sigma(r)) & \text{if } r_1 < r < r_2, \\ 1 & \text{if } r > r_2, \end{cases}$$

where

$$s(r) = \int_{r_1}^{r} (t - r_1)^2 (t - r_2)^2 \, dt.$$

Set $\gamma = 1 + i\sigma$ and $\tilde\gamma = 1 + i\tilde\sigma$. Then, with the coefficient matrices set to

$$\mathcal{A}(\boldsymbol{x}) = \frac{1}{r^2} \begin{pmatrix} \gamma x^2 + \tilde\gamma y^2 & xy(\gamma - \tilde\gamma) \\ xy(\gamma - \tilde\gamma) & \gamma y^2 + \tilde\gamma x^2 \end{pmatrix}, \qquad \mathcal{B}(\boldsymbol{x}) = \varepsilon(\boldsymbol{x})\gamma\tilde\gamma \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$



(a) The case $r_3 = 3$. (b) The case $r_3 = 20$.

FIG. 1. *Finite element meshes (the dielectric is in darker shade).*

the weak formulation of the truncated PML resonance problem is to find eigenvalues $\lambda \equiv k^2$ satisfying

$$\langle \mathcal{A}\nabla u, \nabla v\rangle = \lambda \langle \mathcal{B}u, v\rangle \quad \text{for all } v \in H_0^1(B_{r_3}) \tag{3.10}$$

for some nontrivial eigenfunction u in $H_0^1(B_{r_3})$. Here $B_{r_3} = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| < r_3\}$, and $\langle\cdot,\cdot\rangle$ denotes the $L^2(B_{r_3})$ inner product. To discretize (3.10), we used Lagrange finite elements of degree one on the meshes shown in Figure 1(a) and 1(b). The two meshes correspond to the two values of r_3 that we will investigate.

The discretization results in a large sparse generalized eigenvalue problem

$$\mathsf{A}\mathsf{x} = \lambda\, \mathsf{B}\mathsf{x}, \tag{3.11}$$

where $\mathsf{A}_{ij} = \langle \mathcal{A}\nabla\phi_j, \nabla\phi_j\rangle$, $\mathsf{B}_{ij} = \langle \mathcal{B}\phi_j, \phi_i\rangle$, and ϕ_i's are the usual nodal finite element basis. Note that this is a standard linear eigenvalue problem, because we have used PML coefficients that do *not* depend on frequency (unlike the ones in the original paper of Berenger [1]). We shall see that in the asymptotic integral equation approach in the next subsection, we will get a nonlinear eigenproblem. Now, recall the exact resonant k values calculated in Table 1. We compare the square roots of the eigenvalues computed by (3.11) with the exact values of Table 1.

The square root of the full spectrum of (3.11) in the case of the mesh in Figure 1(a) (the $r_3 = 3$ case), together with the first few exact resonance values, is shown in Figure 2. As marked in the figure, a number of points in the computed spectrum lie far away from the exact resonances and must clearly be considered spurious.

Next, we systematically investigate the convergence of the first ten nonspurious resonance values to the exact value with respect to the discretization meshsize. Starting with the mesh in Figure 1(a), we perform a series of successive refinements. The mesh at refinement level J is obtained by joining the midpoint of the edges of each triangle of the previous level $J-1$. At each refinement, the coordinates of the newly created vertices on the dielectric-air interface are adjusted so that they lie exactly on the unit circle. Let us denote by k_ℓ^J the ℓth resonance value computed at refinement level J, where the ordering in ℓ is with respect to increasing real part, considering

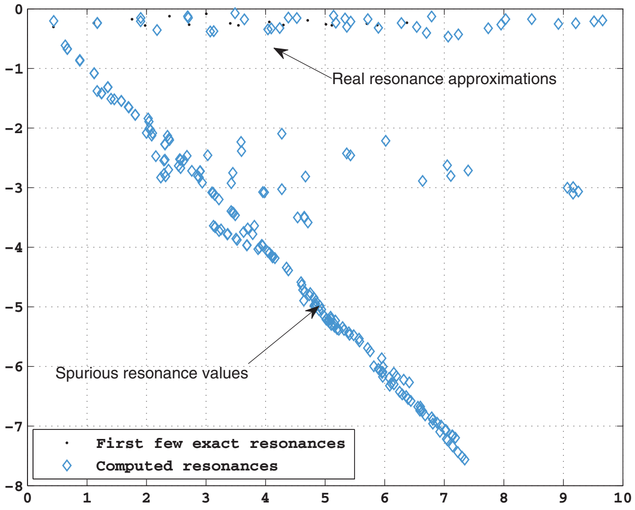


FIG. 2. *Computed resonances for the case* $r_3 = 3$.

only the nonspurious values. The results are tabulated in Table 2. The definitions of mean orders of convergence in the table are as follows:

$$\text{apparent mean order of convergence for } \ell\text{th resonance} := \frac{1}{3}\sum_{J=2}^{4}\log_2\left(\frac{|k_\ell^J - k_\ell^{J-1}|}{|k_\ell^J - k_\ell^{J+1}|}\right);$$

$$\text{actual mean order of convergence for } \ell\text{th resonance} := \frac{1}{3}\sum_{J=2}^{4}\log_2\left(\frac{e_\ell^J}{e_\ell^{J+1}}\right),$$

where for each ℓ, the true error e_ℓ^J is defined by $e_\ell^J = |k_\ell^J - k_{n,m}|$ for the n, m values indicated in the first column of the table (under the title "Compare with $k_{n,m}$"). The "apparent" rate quantifies the order of difference of approximations from successive refinements and is a standard way to measure convergence rate in cases where we have no knowledge of the exact solution.

It is important to note the difference between the apparent and actual rates of convergence in the case of some resonance values (see rows with $\ell = 1, 2, 3$ in Table 2). These suggest that although the computed resonances appear to converge at a second order rate, they converge to the wrong limit. We conjecture that this is due to the spectral changes caused by the truncation of PML at radius r_3. Consider the results in Table 3, where we report the resonance values obtained using the mesh in Figure 1(b), with $r_3 = 20$. Clearly there is a marked improvement in the actual convergence rates, supporting the conjecture that r_3 needs to be sufficiently large.

To summarize, we note the following difficulties encountered with the PML approach:

- It is necessary to separate the true eigenvalues from the spurious eigenvalues.
- Although eigenvalues may appear to converge, they can converge to the wrong value if the domain is not large enough.

These problems were readily identified in this validation experiment because we know the exact solution. However, in a situation without any a priori knowledge of the exact solution, it is important to keep in mind that such difficulties can occur.

Table 2
Apparent and actual convergence of computed resonance values.

Case $r_3 = 3$					
$\ell \backslash J$	Computed resonance approximations k_ℓ^J (displayed up to 3 digits)				
	Level 1	Level 2	Level 3	Level 4	Level 5
1	$0.441 - 0.202i$	$0.432 - 0.207i$	$0.430 - 0.208i$	$0.429 - 0.209i$	$0.429 - 0.209i$
2	$1.173 - 0.239i$	$1.129 - 0.240i$	$1.118 - 0.241i$	$1.115 - 0.241i$	$1.114 - 0.241i$
3	$1.175 - 0.236i$	$1.129 - 0.239i$	$1.118 - 0.241i$	$1.115 - 0.241i$	$1.114 - 0.241i$
4	$1.901 - 0.209i$	$1.793 - 0.181i$	$1.765 - 0.176i$	$1.759 - 0.175i$	$1.757 - 0.174i$
5	$1.902 - 0.151i$	$1.793 - 0.170i$	$1.766 - 0.173i$	$1.759 - 0.174i$	$1.757 - 0.174i$
6	$2.176 - 0.356i$	$2.025 - 0.296i$	$1.990 - 0.283i$	$1.981 - 0.280i$	$1.978 - 0.279i$
7	$2.692 - 0.125i$	$2.463 - 0.123i$	$2.404 - 0.122i$	$2.389 - 0.122i$	$2.385 - 0.122i$
8	$2.700 - 0.150i$	$2.464 - 0.128i$	$2.404 - 0.123i$	$2.389 - 0.122i$	$2.385 - 0.122i$
9	$3.070 - 0.380i$	$2.814 - 0.294i$	$2.742 - 0.273i$	$2.723 - 0.268i$	$2.718 - 0.267i$
10	$3.128 - 0.374i$	$2.822 - 0.295i$	$2.744 - 0.274i$	$2.724 - 0.268i$	$2.719 - 0.267i$
	Apparent mean order	Difference between resonances from successive refinements			
ℓ	of convergence	$\lvert k_\ell^1 - k_\ell^2\rvert$	$\lvert k_\ell^2 - k_\ell^3\rvert$	$\lvert k_\ell^3 - k_\ell^4\rvert$	$\lvert k_\ell^4 - k_\ell^5\rvert$
1	1.90	0.0100	0.0029	0.0007	0.0002
2	1.97	0.0440	0.0114	0.0029	0.0007
3	1.96	0.0452	0.0118	0.0030	0.0008
4	2.00	0.1119	0.0277	0.0070	0.0018
5	1.99	0.1105	0.0276	0.0070	0.0018
6	2.02	0.1628	0.0372	0.0096	0.0024
7	1.97	0.2282	0.0592	0.0151	0.0038
8	1.98	0.2372	0.0597	0.0153	0.0039
9	1.92	0.2698	0.0744	0.0197	0.0050
10	1.96	0.3160	0.0806	0.0212	0.0054
Compare	Actual mean order	Actual errors			
with $k_{n,m}$	of convergence	$J = 2$	$J = 3$	$J = 4$	$J = 5$
$\lvert k_1^J - k_{0,1}\rvert$	0.01	0.0971	0.0957	0.0954	0.0953
$\lvert k_2^J - k_{1,1}\rvert$	0.90	0.0134	0.0024	0.0015	0.0021
$\lvert k_3^J - k_{1,1}\rvert$	0.92	0.0139	0.0024	0.0015	0.0020
$\lvert k_4^J - k_{2,1}\rvert$	1.98	0.0370	0.0093	0.0024	0.0006
$\lvert k_5^J - k_{2,1}\rvert$	1.99	0.0370	0.0093	0.0023	0.0006
$\lvert k_6^J - k_{0,2}\rvert$	1.98	0.0501	0.0128	0.0033	0.0008
$\lvert k_7^J - k_{3,1}\rvert$	1.98	0.0793	0.0201	0.0051	0.0013
$\lvert k_8^J - k_{3,1}\rvert$	1.98	0.0802	0.0205	0.0052	0.0013
$\lvert k_9^J - k_{1,2}\rvert$	1.96	0.1009	0.0264	0.0067	0.0017
$\lvert k_{10}^J - k_{1,2}\rvert$	1.96	0.1090	0.0285	0.0073	0.0018
Degrees of freedom:	233	969	3,953	15,969	64,193

For instance, in our experiments with PML in cases when an exact solution is unknown (reported in later subsections), we needed to separate the true eigenvalues from the spurious eigenvalues. To identify the true eigenvalues, we used the following techniques: (i) We compared the variations in the computed spectrum when PML parameters were varied. (ii) We also compared the spectrum computed with the standard rectangular (tensor product–type [3]) PML with the results from the circular PML in (3.9). The spectral points that persisted across these changes were considered to be the real eigenvalues. (We shall not report these details here for the sake

Table 3
Convergence when the larger domain is used; cf. Table 2.

Case $r_3 = 20$						
Compare	Actual errors					Mean order
with $k_{n,m}$	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$	of convergence
$\lvert k_1^J - k_{0,1}\rvert$	0.1096	0.0342	0.0105	0.0034	0.0020	1.44
$\lvert k_2^J - k_{1,1}\rvert$	0.1283	0.0411	0.0122	0.0032	0.0008	1.83
$\lvert k_3^J - k_{1,1}\rvert$	0.1299	0.0412	0.0122	0.0032	0.0008	1.83
$\lvert k_4^J - k_{2,1}\rvert$	0.1763	0.0497	0.0140	0.0036	0.0009	1.90
$\lvert k_5^J - k_{2,1}\rvert$	0.1961	0.0545	0.0155	0.0040	0.0010	1.90
$\lvert k_6^J - k_{0,2}\rvert$	0.1376	0.0487	0.0130	0.0033	0.0008	1.84
$\lvert k_7^J - k_{3,1}\rvert$	0.2965	0.0759	0.0201	0.0051	0.0013	1.97
$\lvert k_8^J - k_{3,1}\rvert$	0.3089	0.0781	0.0205	0.0052	0.0013	1.97
$\lvert k_9^J - k_{1,2}\rvert$	0.2237	0.0756	0.0182	0.0046	0.0012	1.89
$\lvert k_{10}^J - k_{1,2}\rvert$	0.2229	0.0762	0.0186	0.0047	0.0012	1.89
Degrees of freedom:	427	1,739	7,021	28,217	113,137	

of brevity.) It is more difficult to overcome the discrepancy between the apparent and actual convergence. We typically experiment with an increasing set of r_3 values, holding meshsize (approximately) fixed, until the variation in the eigenvalues of interest becomes negligible. This often requires meshes with a large number of degrees of freedom and hence entails expensive computations.

3.2. A homogeneous thin membrane. In this subsection, we will describe the integral equation approach to the computation of resonances and compare the results from it to those obtained with PML. The resonating object is a thin homogeneous dielectric membrane occupying the rectangular region $[-0.5, 0.5] \times [-h/2, h/2]$. The dielectric constant is set to the following function:

$$\varepsilon(x, z) = \begin{cases} 6/h & \text{if } |z| < h/2, \quad x \in [-0.5, 0.5], \\ 1 & \text{otherwise;} \end{cases} \tag{3.12}$$

i.e., we choose $\varepsilon_0(x) \equiv 6$. For our numerical experiments here, we choose a geometrically decreasing sequence of values for h.

Let us first describe the collocation discretization of the asymptotic integral equation derived in section 2. The computational domain, which is now $[-0.5, 0.5]$, is meshed by a grid of evenly spaced points set at a distance δ apart. Discrete approximations to resonance modes are now in the space V_δ of continuous functions which are linear in between adjacent grid points. Define the matrix-valued function $\mathsf{S} : \mathbb{C} \mapsto \mathbb{C}^{N\times N}$ by $\mathsf{S}(\lambda) = \mathsf{I} - \lambda\,\mathsf{T}(\lambda)$, where I denotes the identity matrix, and the entries of the matrix $\mathsf{T}(\lambda)$ are defined by

$$[\mathsf{T}(\lambda)]_{ij} = -\int_{-1/2}^{1/2} \varepsilon_0(x')\, G_\lambda(x_i, 0, x', 0)\, \psi_j(x') \; dx'. \tag{3.13}$$

Here ψ_j is the unique function in V_δ, which is one at the jth grid point and zero at all other grid points. With these notations, the discrete problem is the dense *nonlinear eigenvalue problem* of finding complex numbers λ and corresponding nontrivial vectors x satisfying

$$\mathsf{S}(\lambda)\,\mathsf{x} = 0. \tag{3.14}$$

This can be rewritten as a nonlinear system for λ and x, to which Newton's method or its variants can be applied. To compute the matrix entries defined by (3.13) we split the integral into integrals over each mesh interval (of length δ). On those intervals where the integrand is smooth, the integrals are approximated by high order Gaussian quadratures. We must be more careful in the intervals containing the singularity of G_λ. On such elements, we use an expansion of the integrand to approximate the integral. In all cases, our integral approximations are at least $O(\delta^7)$ accurate.

To solve (3.14), we use the residual inverse iteration analyzed by Neumaier [15].

Algorithm 3.1 (residual inverse iterations).

```
1. Input an initial approximation $\lambda_0$ close to the eigenvalue of interest. If $\mathsf{S}(\lambda_0)$ is invertible, continue.
2. Set (Wilkinson) initial guess $\mathsf{x}_0$ for the eigenvector:
   (a) Perform the QR-factorization $\mathsf{QR} = \mathsf{S}(\lambda_0)$.
   (b) Let $\mathsf{b} = \mathsf{Q}\,\mathsf{n}$, where $\mathsf{n}$ is the vector whose components are 1.
   (c) Solve the linear system $\mathsf{S}(\lambda_0)\tilde{\mathsf{x}}_0 = \mathsf{b}$, by $\tilde{\mathsf{x}}_0 = \mathsf{R}^{-1}\mathsf{n}$.
   (d) Normalize by $\mathsf{x}_0 = \tilde{\mathsf{x}}_0/\mathsf{e}^*\tilde{\mathsf{x}}_0$, where $\mathsf{e}$ is the unit vector with one in the position of the largest entry of $\tilde{\mathsf{x}}_0$.
   (e) Set $\mathsf{y}^* = \mathsf{e}^*\mathsf{R}^{-1}\mathsf{Q}^*$ for use later.
3. For $l = 0, 1, 2, \ldots$ (until a stopping criteria is met) do:
   (a) $\lambda_{l+1} = \lambda_l - \frac{\mathsf{y}^*\mathsf{S}(\lambda_l)\mathsf{x}_l}{\mathsf{y}^*\mathsf{S}'(\lambda_l)\mathsf{x}_l}$, where $[\mathsf{S}'(z)]_{ij} = \frac{d}{dz}[\mathsf{S}(z)]_{ij}$.
   (b) $\tilde{\mathsf{x}}_{l+1} = \mathsf{x}_l - \mathsf{R}^{-1}\mathsf{Q}^*\mathsf{S}(\lambda_{l+1})\mathsf{x}_l$.
   (c) Normalize by $\mathsf{x}_{l+1} = \tilde{\mathsf{x}}_{l+1}/\|\tilde{\mathsf{x}}_{l+1}\|_2$.
```

These iterations can be stopped once $|\lambda_{l+1} - \lambda_l|$ is smaller than a prescribed tolerance. In step 3(a) of the algorithm, we have used one step of a one-dimensional Newton iteration. We can substitute this step with multiple Newton iterations or other nonlinear solvers.

This algorithm works well in our application *if* good initial approximations λ_0 are given. In order to find good initial guesses, we borrowed a technique used for plotting the pseudospectra [18] of matrices. Namely, if $\sigma_{\min}(\mathsf{S})$ denotes the smallest singular value of S, then it is easy to see that

$$\sigma_{\min}(\mathsf{S}(\lambda)) < \delta \quad \text{if and only if} \quad \|\mathsf{S}(\lambda)^{-1}\|_2 > 1/\delta. \tag{3.15}$$

Motivated by this, before launching the residual inverse iterations for fine meshes, we first use a coarse mesh to obtain an inexpensive matrix approximation $\mathsf{S}(\lambda)$. We then compute the minimum singular value $\sigma_{\min}(\mathsf{S}(\lambda))$ on a grid of λ in the complex plane. For coarse meshes, $\mathsf{S}(\lambda)$ is a small matrix, so this computation is fast. Because of (3.15), the plot of the minimum singular values locates regions where the resolvent $\mathsf{S}(\lambda)$ is nearly singular, thus providing good initial guesses for Algorithm 3.1. For our current example of the homogeneous thin membrane, a coarse mesh resulting in a 20×20 matrix function $\mathsf{S}(\lambda)$ yields the contour plot of $\sigma_{\min}$ shown in Figure 3. Note that since $k = \sqrt{\lambda}$ is what we shall report, Figure 3 shows $\sigma_{\min}$ as a function of k (rather than λ).

Next, we report the first few resonance values computed using Algorithm 3.1 applied to (3.14). We mesh the interval $[-0.5, 0.5]$ uniformly with a mesh of meshsize $\delta_0 = 1/20$. To perform a study of convergence with respect to meshsize, we refine this coarse mesh by splitting each grid element into two equal elements, so the meshsize at the refinement level J is $\delta_J = 2^{-(J-1)}/20$. Denoting the ℓth resonance value computed using the mesh at refinement level J as $k_\ell^{J,\infty}$, the differences in the computed resonance values at successive refinements are collected in Table 4. Examining these differences,

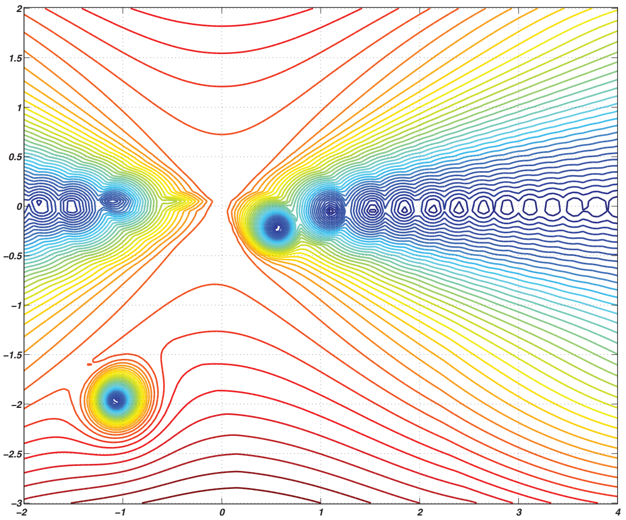


FIG. 3. *Contour plot indicating the resonant frequencies in the k plane.*

TABLE 4
Resonance values (square roots $k = \sqrt{\lambda}$) from the collocation discretization of the asymptotic integral equation.

ℓ	$\lvert k_\ell^{1,\infty} - k_\ell^{2,\infty}\rvert$	$\lvert k_\ell^{2,\infty} - k_\ell^{3,\infty}\rvert$	$\lvert k_\ell^{3,\infty} - k_\ell^{4,\infty}\rvert$	$\lvert k_\ell^{4,\infty} - k_\ell^{5,\infty}\rvert$	$\lvert k_\ell^{5,\infty} - k_\ell^{6,\infty}\rvert$	$\lvert k_\ell^{6,\infty} - k_\ell^{7,\infty}\rvert$	$k_\ell^{*,\infty} := k_\ell^{7,\infty}$
1	0.0002044	0.0000561	0.0000153	0.0000041	0.0000011	0.0000003	$0.5650936 - 0.2220208i$
2	0.0020219	0.0005452	0.0001460	0.0000389	0.0000103	0.0000027	$1.1080760 - 0.0456433i$
3	0.0072890	0.0019383	0.0005112	0.0001341	0.0000350	0.0000091	$1.5159978 - 0.0284500i$
4	0.0163478	0.0043339	0.0011364	0.0002959	0.0000767	0.0000198	$1.8255240 - 0.0175582i$
5	0.0303687	0.0080609	0.0021085	0.0005468	0.0001411	0.0000363	$2.0959593 - 0.0137123i$
6	0.0490834	0.0130928	0.0034242	0.0008862	0.0002280	0.0000584	$2.3307014 - 0.0110730i$
7	0.0733213	0.1801607	0.0073143	0.0018911	0.0004850	0.0001239	$2.7440293 - 0.0080844i$
8	0.0964783	0.0377803	0.0099407	0.0025712	0.0006591	0.0001681	$2.9294654 - 0.0065552i$
9	0.0523806	0.0493936	0.0130412	0.0033757	0.0008650	0.0002205	$3.1026198 - 0.0063434i$
10	0.0090552	0.0628079	0.0166530	0.0043152	0.0011058	0.0002817	$3.2674525 - 0.0052055i$

we conjecture that the convergence rate of collocation discretization for the resonance values is $O(\delta^2)$, where δ is the meshsize. The last column of Table 4 lists the resonance values computed using the finest mesh.

We now compare these resonance values with those obtained using the PML approach. We enclose the dielectric in $[-0.5, 0.5] \times [-h/2, h/2]$ by circles of radius r_1, r_2, and r_3 and set the PML parameters as described in section 3.1. We experimented with a number of r_3 values and concluded that selecting $r_3 = 20$ seems appropriate to get good approximations in this example. We also needed to isolate the spurious eigenvalues from the interesting ones (see remarks at the end of section 3.1). We shall consider the following geometrically decreasing sequence of membrane thicknesses:

$$h = \frac{0.25}{2^{L-1}}, \qquad L = 1, 2, \ldots, 7.$$

For each L value, we mesh the domain (B_{r_3}) such that mesh aligns with the dielectric boundaries. Furthermore, the meshes are such that for all values of h considered, the dielectric region will always have four layers of elements. The meshsize inside the dielectric is thus maintained approximately at $h/4$ by constraining the angles of

Table 5
Difference between the asymptotic and PML resonance approximations.

ℓ	$\lvert k_\ell^{\text{pml},1} - k_\ell^{*,\infty}\rvert$	$\lvert k_\ell^{\text{pml},2} - k_\ell^{*,\infty}\rvert$	$\lvert k_\ell^{\text{pml},3} - k_\ell^{*,\infty}\rvert$	$\lvert k_\ell^{\text{pml},4} - k_\ell^{*,\infty}\rvert$	$\lvert k_\ell^{\text{pml},5} - k_\ell^{*,\infty}\rvert$
1	0.0411	0.0205	0.0101	0.0048	0.0021
2	0.1655	0.0864	0.0444	0.0226	0.0115
3	0.4086	0.2144	0.1102	0.0561	0.0284
4	0.7001	0.3684	0.1896	0.0965	0.0488
5	0.6427	0.5498	0.2841	0.1448	0.0733
6	0.7790	0.7500	0.3884	0.1981	0.1003

the mesh triangles to never deteriorate below 25 degrees. With $\varepsilon(x, y)$ as in (3.12), we then solve the resulting finite element eigenproblem (3.11) for the first few eigenvalues and compare them with the resonance approximations previously displayed in the last column of Table 4.

One of our aims in this comparison is the verification of the theoretically predicted asymptotic convergence rate of $O(h)$ of Theorem 2.4. To realize this goal, we must avoid discretization errors as much as we can, but without going to prohibitively expensive meshsizes. Note that the first six resonance values in Table 4 have stabilized up to four digits at the seventh level of refinement, so we denote these six values by $k_\ell^{*,\infty}$, $\ell = 1, 2, \ldots, 6$, and use them for the comparison with the corresponding first six resonance approximations from PML. In order to avoid finite element discretization errors in the comparable PML resonance approximations, we perform multiple refinements of the finite element mesh until their first six resonance approximations have no variation in at least the first two significant digits. Denoting these approximations by $k_\ell^{\text{pml},L}$ for the case of membrane thickness $h = 0.25/2^{L-1}$, we display in Table 5 the distance of these approximations to the asymptotic ones. The linear asymptotic convergence rate is clearly apparent.

Next we apply Corollary 2.5 to attempt to improve the asymptotic resonance approximations from the problem in the previous subsection. Recall that from the residual correction procedure described there, we have limiting resonance value λ_0 and a discrete approximation to the corresponding resonance function $u_0(x)$. We will apply these values to calculate

$$\lambda_0 + h\lambda_1$$

to get what should be a better approximation to the resonance value λ_h for a given total slab thickness h. Note that the correction

$$\lambda_1 = \lambda_0^2 \frac{\langle g, u_0\rangle}{1 + \lambda_0^2 \langle DT_0(\lambda_0)u_0, u_0\rangle},$$

where

$$g(x) = -\int_\Omega G_{\lambda_0}(x, 0, x', 0)u_0(x')dx' + \frac{\varepsilon_0(x)u_0(x)}{2}\left(\zeta^2 + \frac{1}{4}\right),$$

involves merely double integrations over Ω, in this case a one-dimensional domain. In the second term in the numerator the integration in ζ can be calculated exactly. The integral in the denominator is

$$\langle DT_0(\lambda_0)u_0, u_0\rangle = -\int_{-.5}^{.5}\int_{-.5}^{.5} \frac{i\varepsilon_0(x')}{8\sqrt{\lambda_0}} H_1^{(1)}\left(\sqrt{\lambda_0}|x - x'|\right)|x - x'|u_0(x')\overline{u_0}(x)dx'dx.$$

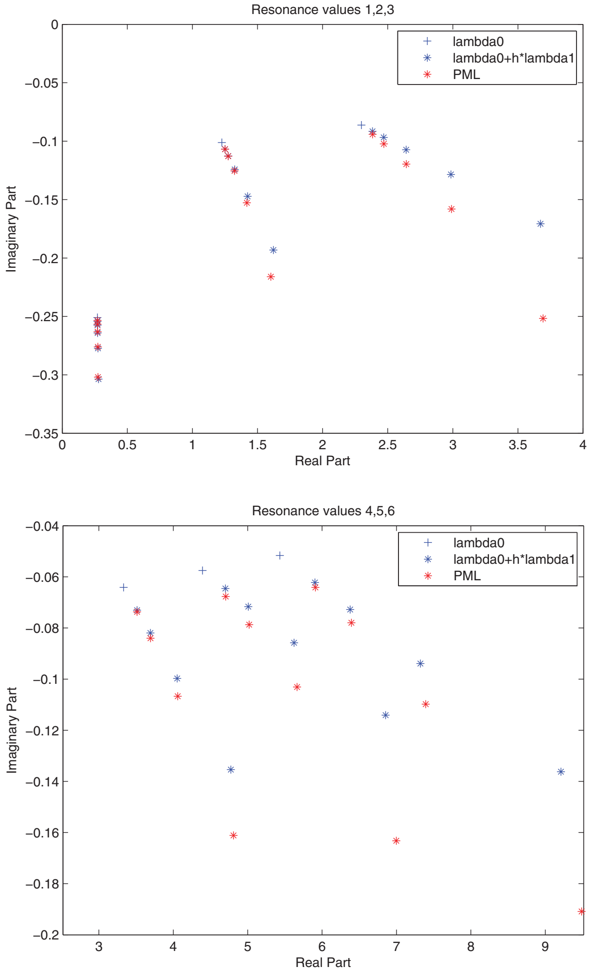


FIG. 4. *The first six computed resonance values λ in the complex plane for varying slab thicknesses.*

For both of the double integrals, we compute the inner integral using the piecewise linear basis functions for u_0, and for the outer integral we use the trapezoid rule. In all computations that follow the meshsize was $\delta = 1/640$, the sixth refinement level, for which we believe the calculations of the first six limiting resonance values λ_0 are accurate up to about four significant digits. Recall that the PML values used from the previous section are accurate to about two significant digits.

Figure 4 shows all of the computed values plotted on the complex plane, and Figure 5 gives a log-log plot of the errors. For the first resonance, the corrected asymptotic values are within the presumed accuracy of the PML approximation for all values of h, and hence we see the convergence flattening in the log-log plot. For the third resonance, the convergence appears only slightly more than linear, but all of the other values exhibit the significantly better than linear convergence expected from Corollary 2.5.

Remark 3.1. The approximation from [14] used to obtain Corollary 2.5 deteriorates for larger frequencies, and we therefore expect that for the higher-numbered

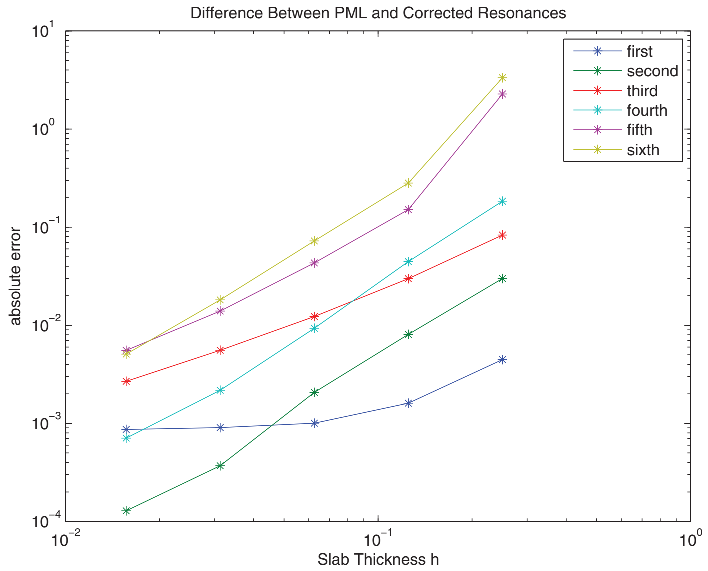


FIG. 5. *Log-log plot of the absolute error between the PML computed and corrected asymptotic resonance values.*

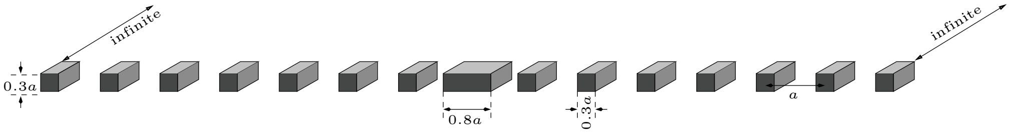


FIG. 6. *The photonic dielectric structure.*

resonances, the formula from Theorem 2.4 will be far superior. The computation of this more accurate correction will involve computing the application of the higher dimensional integral operator T_h, but will not require any inversion.

3.3. A photonic membrane. In this subsection, we will describe the results of computation from a thin photonic membrane having a periodic dielectric pattern with a defect. The structure is shown in Figure 6 and was previously investigated in [5] by time domain methods. We will give results from both the PML and the asymptotic integral equation approaches. Unlike the previous subsections, our purpose here is not a convergence study, but rather a comparison with the results in [5]. The structure in Figure 6 is invariant in the third direction, so the model is reduced to one in the xz plane perpendicular to the symmetry direction. Note that while fully periodic structures have band gaps, this structure is not periodic in that it has a defect in the center and has finite extent in the plane. Hence instead of defect eigenvalues, we seek resonances.

We choose the dielectric constant as in [5]; namely, in the central defect column and the fourteen off-center columns, $\varepsilon(\boldsymbol{x})$ is 13, while in the remaining region it equals 1. We set the scaling parameter a in Figure 6 to $1/14.3$ so that the xz cross section fits into $[-0.5, 0.5] \times [-h/2, h/2]$ with $h = 0.3a$. Since this h is small, it is reasonable to attempt the asymptotic approach.

For the PML calculations, we enclose the xz cross section of the photonic structure by disks of radius $r_1 = 0.6$, $r_2 = 2$, and $r_3 = 10$ and set the PML parameters as in the previous sections. This domain is meshed such that there are at least four layers

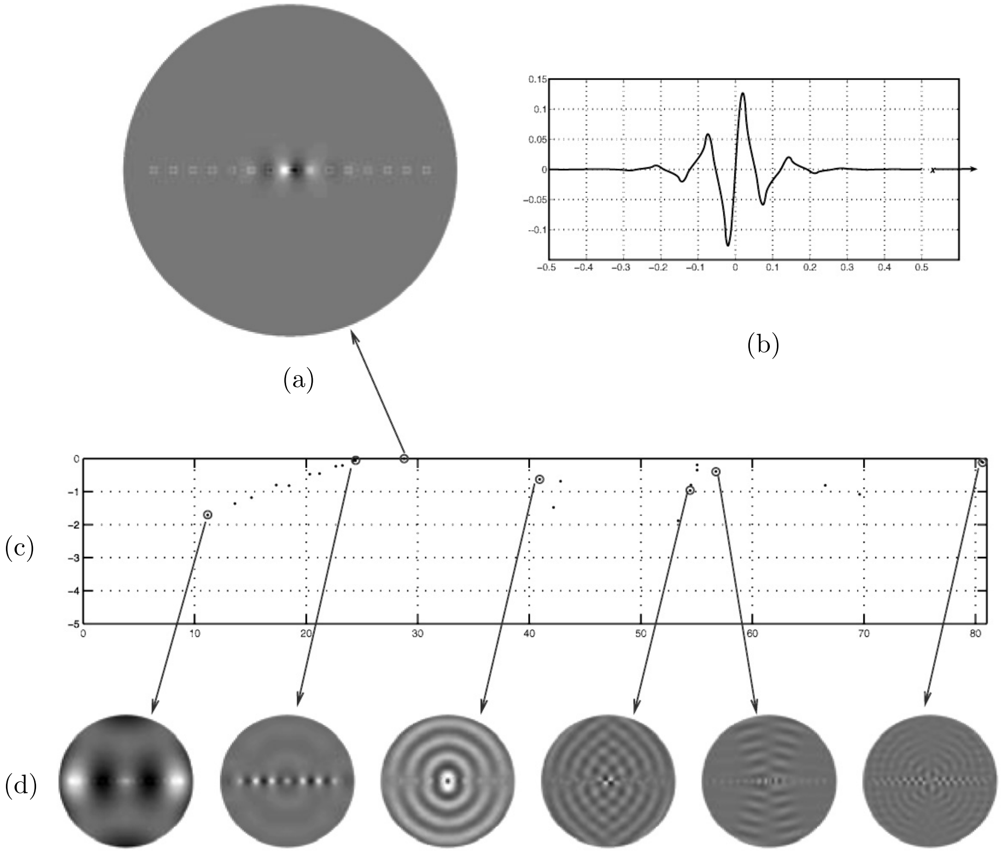


Fig. 7. *A few resonances of the photonic membrane.* (a) *The highly localized resonance mode corresponding to $k \approx 28.7878 - 0.0017i$ obtained using PML.* (b) *The corresponding mode obtained using the asymptotic approach, plotted on the limiting domain* $[-0.5, 0.5]$. (c) *Point plot of a few resonant k values computed using PML.* (d) *The modes corresponding to the k values circled in* (c). *(In all plots of the resonance modes, only the real part is plotted and only the region $r < r_1$ is shown.)*

of elements across the membrane thickness (and the mesh coarsens away from the dielectric). The computations then proceed similarly, except that to obtain higher accuracy, we now use Lagrange finite elements of degree five. The results are shown in Figure 7. The mode shown in Figure 7(a) is qualitatively similar to the one in [5, Fig. 5]. Furthermore, its corresponding resonance value is such that $ka/2\pi \approx 0.321$, a number close to the frequency of 0.313 reported in [5]. As seen from Figure 7(a), this mode is highly localized near the defect. Although there are many other resonances, as seen from Figure 7(d), localization near the defect or even near the membrane seems to be uncommon. (One other mode that is somewhat localized within the membrane is seen in the second plot of Figure 7(d).)

For the asymptotic approach, we set $\varepsilon_0 = \varepsilon h$ with the ε and h as described above and solve the resulting one-dimensional nonlinear eigenproblem on $\Omega = [-0.5, 0.5]$. The initial guesses for the nonlinear eigenvalue solver were obtained using the pseudospectrum-like plot in Figure 8 (computed as described previously; see Figure 3).

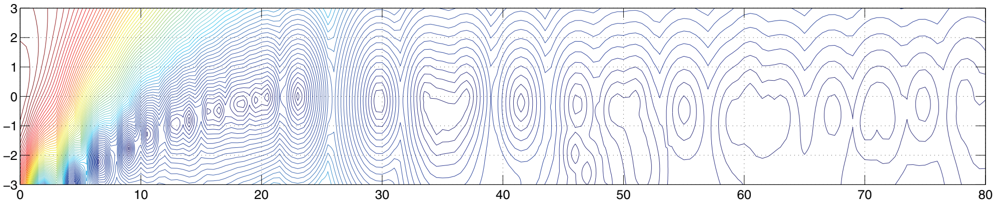


FIG. 8. *Contour plot in the complex k plane indicating the locations of the asymptotic resonances.*

TABLE 6
A few values of resonances k for the photonic membrane.

Values from PML	Asymptotic values	Corrected asymptotic values
$20.3154 - 0.4704i$	$17.8719 - 0.2655i$	$19.9420 - 0.4049i$
$22.6512 - 0.2292i$	$19.7458 - 0.1116i$	$22.7624 - 0.1228i$
$24.2375 - 0.0658i$	$20.3675 - 0.0348i$	$23.7018 - 0.0493i$
$28.7878 - 0.0017i$	$23.0690 - 0.0006i$	$28.0236 - 0.0005i$
$40.9258 - 0.6283i$	$29.7120 - 0.1592i$	$39.1908 - 0.4074i$

A few resonance values obtained using Algorithm 3.1 are reported in Table 6. The table also gives the corresponding resonance approximations from the PML calculations. There is good agreement between the PML and asymptotic values, especially after the correction. The most interesting mode is of course the one localized in the defect. The limiting, uncorrected resonance value for this mode is such that $ka/2\pi \approx 0.257$, and produced the nonlinear eigenfunction in Figure 7(b). This qualitatively resembles not only the mode plot in [5, Fig. 5], but also the trace of the corresponding mode obtained from PML (Figure 7(a)) on the x-axis (the centerline of the dielectric). The corrected asymptotic resonance value is such that $ka/2\pi \approx 0.311$, very close to the value 0.313 reported in [5].

Considering that the PML eigenvalue problem we solved is of size 221201×221201, while the asymptotic problem is only of size 2289×2289, the advantages of the asymptotic approach are clearly evident for this particular geometry.

4. Discussion. We propose two methods for calculating resonance associated with the scalar wave equation. The first method is suited for thin, high index structures which are gaining popularity in the photonic band gap community. It is an asymptotic method that exploits the specifics of the problem and allows for the calculation of resonance to be carried out in one dimension less than the spatial dimension of the problem. The second method, based on the finite element method with the PML, is a general approach which is not restricted to thin structures. In this work, we examine the convergence properties of the finite element approach and use it to verify the approximation properties of the asymptotic method. A final set of calculations with both methods reproduces a photonic band gap resonance calculation reported in the literature.

For a thin membrane structure with high index, we find that the asymptotic method (2.3) is particularly effective. When discretized, it leads to a dense, but small, nonlinear eigenvalue problem. While we established approximation properties of the method, the numerical evidence is quite convincing. A higher order correction, which is easy to implement, provides more accuracy at a low cost. In comparison with the PML approach, the asymptotic method has the clear advantage of a smaller system, brought about by the dimensional reduction. Its disadvantage lies in the complications involved in solving a dense nonlinear eigenproblem.

The finite element PML approach is attractive because the matrices involved are sparse and the eigenvalue problem to find resonance is linear. It is also more widely applicable. One challenge in using this method is the presence of spurious modes. Our experience is that it is possible to identify spurious modes. Another unattractive feature of the PML approach is that the resonance values may appear to converge under refinement, but to incorrect limits. Our experience with this method, while limited, does gives us hope that it is possible to deal with these difficulties, and that it is possible to develop a robust finite element–based method for calculating resonance. It remains to be seen, however, if it is a viable alternative to simple FDTD calculations.

Acknowledgment. The authors thank Joseph Pasciak for several enlightening discussions on exterior eigenvalue approximations.

REFERENCES

[1] J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[2] J. Bramble and J. Pasciak, *Analysis of a Finite Element PML approximation for the three dimensional time-harmonic Maxwell problem*, Math. Comp., 77 (2008), pp. 1–10.

[3] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.

[4] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1992.

[5] S. H. Fan, J. N. Winn, A. Devenyi, J. C. Chen, R. D. Meade, and J. D. Joannopoulos, *Guided and defect modes in periodic dielectric waveguides*, J. Opt. Soc. Amer. Ser. B, 12 (1995), pp. 1267–1272.

[6] A. Figotin and A. Klein, *Localization of light in lossless inhomogeneous dielectrics*, J. Opt. Soc. Amer. Ser. A, 15 (1998), pp. 1423–1435.

[7] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[8] R. Froese, *personal communication*.

[9] S. Hein, T. Hohage, W. Koch, and J. Schöberl, *Acoustic resonances in a high-lift configuration*, J. Fluid Mech., 582 (2007), pp. 179–202.

[10] T. Kato, *Perturbation Theory for Linear Operators*, Classics in Math., Springer-Verlag, Berlin, 1995; reprint of the 1980 edition.

[11] W. Koch, *Acoustic resonances in rectangular open cavities*, AIAA J., 43 (2005), pp. 2345–2349.

[12] K. D. Kokkotas and B. G. Schmidt, *Quasi-normal modes of stars and black holes*, Living Rev. Relativ., 2 (1999); available online at http://www.livingreviews.org/lrr-1999-2.

[13] M. Lenoir, M. Vullierme-Ledard, and C. Hazard, *Variational formulations for the determination of resonant states in scattering problems*, SIAM J. Math. Anal., 23 (1992), pp. 579–608.

[14] S. Moskow, F. Santosa, and J. Zhang, *An approximate method for scattering by thin structures*, SIAM J. Appl. Math., 66 (2005), pp. 187–205.

[15] A. Neumaier, *Residual inverse iteration for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal., 22 (1985), pp. 914–923.

[16] J. E. Osborn, *Spectral approximation for compact operators*, Math. Comp., 29 (1975), pp. 712–725.

[17] M. Rechberger, *Numerical Methods for Simulation of Acoustic Resonances*, Master's thesis, Johannes Kepler Universität Linz, Linz, Austria, 2005.

[18] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.

[19] P. Villeneuve, S. Fan, S. Johnson, and J. Joannopoulos, *Three-dimensional photon confinement in photonic crystals of low-dimensional periodicity*, IEE Proc. Optoelectron., 145 (1998), pp. 384–390.

[20] J. Vuckovic, M. Loncar, H. Mabuchi, and A. Scherer, *Optimization of the q factor in photonic crystal microcavities*, IEEE J. Quantum Elec., 38 (2002), pp. 850–856.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 64–80

A LEVEL SET APPROACH TO ANISOTROPIC SURFACE EVOLUTION WITH FREE ADATOMS*

CHRISTINA STÖCKER† AND AXEL VOIGT‡

Abstract. We variationally derive a thermodynamically consistent model for surface evolution under the influence of free adatoms. The resulting system of nonlinear partial differential equations couples a diffusion equation on a surface to the evolution of the surface. A numerical approach based on a finite element discretization of a level set equation is described for an anisotropic evolution, with nonconvex anisotropy functions. Various simulation examples in two and three dimensions demonstrate the applicability of the method.

Key words. thin film growth, anisotropic evolution, PDEs on surfaces, higher order evolution equations, level set method, finite elements, adaptivity

AMS subject classifications. 35K55, 35R35, 58J32

DOI. 10.1137/060678166

1. Introduction.

Evolving interfaces are a key ingredient in many problems in materials science. As smaller and smaller length scales become of interest, the importance of interfaces compared with phenomena in the bulk phases increases and therefore requires more detailed considerations than in classical, more macroscopic models. To model grain boundaries, solid-vapor interfaces, or coherent phase transitions, various geometric evolution laws have been proposed. Mean curvature flow and surface diffusion are two prominent examples. In its basic form these laws have been derived by Mullins [14, 15]. The equations follow from a surface free energy

$$E[\Gamma] = \int_\Gamma \gamma \, d\Gamma,$$

with $\Gamma = \Gamma(t)$ a compact smooth connected and oriented hypersurface in $\mathbb{R}^{d+1}$ ($d = 1, 2$) and $\gamma = \gamma(\mathbf{n})$ the surface free energy density, possibly depending on the normal to the surface $\mathbf{n}$. The geometric evolution laws are defined as gradient flows of the energy and read, e.g.,

$$V = -kH_\gamma, \tag{1.1}$$

$$bV = -(H_\gamma - c), \tag{1.2}$$

$$V = \nabla_\Gamma \cdot \big(\nu \nabla_\Gamma H_\gamma\big). \tag{1.3}$$

Here $k = k(\mathbf{n})$ is the evaporation modulus, $b = b(\mathbf{n})$ is a kinetic coefficient, $\nu = \nu(\mathbf{n})$ denotes the surface mobility, V is the normal velocity, H_γ is weighted mean curvature

$$H_\gamma = \sum_{i=1}^{d-1} \partial_{p_i p_i} \gamma(\mathbf{n}) \kappa_i,$$

*Received by the editors December 18, 2006; accepted for publication (in revised form) May 5, 2008; published electronically August 1, 2008. This work was supported by EU and NSF through "MagDot" and by DFG through SFB609.

http://www.siam.org/journals/siap/69-1/67816.html

†Institut für Numerische und Angewandte Mathematik, Universität Münster, 48147 Münster, Germany.

‡Institut für Wissenschaftliches Rechnen, Technische Universität Dresden, 01062 Dresden, Germany (axel.voigt@tu-dresden.de).

which is $\frac{\delta E}{\delta \Gamma} \cdot \mathbf{n}$, the variational derivative with respect to variations in Γ in normal direction. Here $\partial_{p_i p_i}$ denotes the second derivative of the one-homogeneous extension of $\gamma : S^d \subset \mathbb{R}^{d+1} \to \mathbb{R}$ with respect to $\mathbf{n}$ in the ith principal direction, and κ_i, $i = 1, \dots, d$, are the principal curvatures. The coefficient $c = \int_\Gamma b^{-1} H_\gamma \, d\Gamma \,/\, \int_\Gamma b^{-1} \, d\Gamma$ is the averaged weighted mean curvature. Equation (1.1) models motion by evaporation-condensation, (1.2) describes the kinetics associated with the rearrangement of atoms on the surface, and (1.3) accounts for diffusion of atoms along the surface. Equation (1.1) is known as curvature flow, (1.2) as volume conserved curvature flow, and (1.3) as motion by surface diffusion. A more general geometric evolution law combining the three effects has been discussed by Fried and Gurtin [9]:

$$V = \nabla_\Gamma \cdot \big(\nu \nabla_\Gamma (H_\gamma + bV)\big) - k(H_\gamma + bV). \tag{1.4}$$

This equation incorporates surface diffusion, evaporation-condensation, and kinetics. We obtain the individual laws discussed above as follows:

- $\nu = 0$, $b = 0 \Rightarrow$ evaporation-condensation (1.1).
- $\nu = \infty$, $k = 0 \Rightarrow$ kinetic (1.2).
- $b = 0$, $k = 0 \Rightarrow$ surface diffusion (1.3).

The special case $k = 0$ has been considered by Cahn and Taylor [6] and can also be written as

$$V = -\big(\nabla_\Gamma \cdot (\nu \nabla_\Gamma)\big) \left(\nabla \cdot (\nu \nabla_\Gamma) - \frac{1}{b} \right)^{-1} \left(\frac{1}{b} H_\gamma \right). \tag{1.5}$$

Another special case is obtained for $\nu = 0$. The resulting equation reads

$$\left(\frac{1}{k} + b \right) V = -H_\gamma, \tag{1.6}$$

which again is a curvature flow equation, but with a modified kinetic coefficient.

More recently Fried and Gurtin [9] introduced a refined model of (1.4) which includes free adatoms. Adatoms are mobile atoms on the surface. They diffuse along the surface and attach and detach from surface defects, which contributes to the evolution of the surface. Especially in the case of solid-vapor interfaces such adatoms are assumed to play an important role in the dynamics of the surface evolution. The model reads

$$\partial_t u + V + uHV = \nabla_\Gamma \cdot (\nu \nabla_\Gamma \mu) - k\mu, \tag{1.7}$$

$$\mu = \partial_u \gamma, \tag{1.8}$$

$$bV + H_{\gamma_1} + \gamma_2 H - \mu - uH\mu = 0, \tag{1.9}$$

with u the adatom density, H the mean curvature, and μ the surface chemical potential, which is defined as the partial derivative of the surface free energy density $\gamma(\mathbf{n}, u) = \gamma_1(\mathbf{n}) + \gamma_2(u)$ with respect to u, and H_{γ_1} again the weighted mean curvature, given as the normal component of the variational derivative of γ_1 with respect to variations in Γ. If u is set to zero, the classic law (1.4) is obtained again with $\gamma = \gamma_1$ and $\mu = H_\gamma + bV$. The model is derived within the framework of configurational forces. Burger [3] used the principle of minimal work to derive the model and analyzed it with $k = 0$ in detail and numerically solved the isotropic case within a graph formulation. A phase-field approximation in this situation is considered by Rätz and Voigt [17].

We will show that the model can also be derived variationally from an energy

$$E[\Gamma, u] = \int_\Gamma \gamma \, d\Gamma,$$

with $\gamma = \gamma(\mathbf{n}, u) = \gamma_1(\mathbf{n}) + \gamma_2(u)$. The more general case, in which the dependency on $\mathbf{n}$ and u cannot be split, will lead to an even more complicated model in which also the adatom energy will be anisotropic. The functional form of such a dependency, however, is not known for any material. We therefore restrict our treatment to the separable case, which leads to the model introduced in [9]. We will introduce a numerical approach based on a level set method and allow for various anisotropies.

The outline of the paper is as follows: In section 2 we derive the model and show its thermodynamic consistency. In section 3 we describe a level set algorithm to solve the adatom surface diffusion model. In section 4 we give some details on the implementation in AMDiS [20] and show numerical results on the evolution towards the equilibrium shape.

2. Model derivation. In order to define a thermodynamically consistent dynamic model for the evolution of the surface we need $\frac{d}{dt}E \leq 0$, with $E = E[\Gamma, u]$. The time derivative of E implies

$$\frac{d}{dt}E = \int_\Gamma \partial_t u \frac{\delta E}{\delta u} \, d\Gamma + \int_\Gamma \mathbf{v} \cdot \frac{\delta E}{\delta \Gamma} \, d\Gamma, \tag{2.1}$$

with $\frac{\delta E}{\delta u}$ the variational derivative of E with respect to u and $\frac{\delta E}{\delta \Gamma}$ the variational derivative of E with respect to Γ, given by

$$\frac{\delta E}{\delta u} = \mu, \tag{2.2}$$

$$\frac{\delta E}{\delta \Gamma} = H_{\gamma_1}\mathbf{n} + \gamma_2 H \mathbf{n}, \tag{2.3}$$

with the chemical potential $\mu = \partial_u \gamma$ and the weighted mean curvature defined above. We decompose the velocity $\mathbf{v}$ into normal and tangential components through $\mathbf{v} = V\mathbf{n} + \mathbf{T}$. This decomposition might be used to obtain the identities

$$\nabla_\Gamma \cdot \mathbf{v} = \nabla_\Gamma \cdot (V\mathbf{n} + \mathbf{T}) = V\nabla_\Gamma \cdot \mathbf{n} + \nabla_\Gamma \cdot \mathbf{T} = VH + \nabla_\Gamma \cdot \mathbf{T}, \tag{2.4}$$

where $H = \nabla_\Gamma \cdot \mathbf{n}$ is used, and

$$\mathbf{v} \cdot \nabla u = (V\mathbf{n} + \mathbf{T}) \cdot \nabla u = V\frac{\partial u}{\partial \mathbf{n}} + \mathbf{T} \cdot \nabla_\Gamma u, \tag{2.5}$$

which will be needed in the following derivation.

A basic conservation law we wish to establish is the invariance of mass in time. The mass in the system is given by

$$m = \int_{\Omega_s} 1 \, d\Omega + \int_\Gamma u \, d\Gamma,$$

with $\Omega_s = \Omega_s(t)$ being the solid domain. Defining an arbitrary portion Σ of Γ and Λ of Ω_s, such that $\Sigma \subset \partial\Lambda$ and ignoring bulk diffusion, we have

$$\frac{d}{dt}m_{|\Sigma} = -\int_{\partial\Sigma} \mathbf{q} \cdot \mathbf{m} \, ds + \int_\Sigma f \, ds, \tag{2.6}$$

with $\partial\Sigma$ the boundary of Σ, $\mathbf{m}$ the conormal on $\partial\Sigma$, $\mathbf{q}$ a tangent surface flux, and f the supply from the vapor. The transport theorems (see, e.g., [1]) imply

$$\begin{aligned} \frac{d}{dt} m_{|\Sigma} = \frac{d}{dt}\left(\int_\Lambda 1\, d\Omega + \int_\Sigma u\, d\Gamma\right) &= \int_\Sigma V\, d\Gamma + \int_\Sigma \dot{u} + u\nabla_\Gamma \cdot \mathbf{v}\, d\Gamma \\ &= \int_\Sigma V + \dot{u} + uVH + u\nabla_\Gamma \cdot \mathbf{T}. \end{aligned} \tag{2.7}$$

$\dot{u}$ denotes the normal time derivative of the interfacial field u (see, e.g., [9])

$$\dot{u} = \partial_t u + \mathbf{T} \cdot \nabla_\Gamma u, \tag{2.8}$$

where we assume that u is constant in a normal direction. Together with the formula for integration by parts on Σ,

$$\int_{\partial\Sigma} \mathbf{q} \cdot \mathbf{m}\, ds = \int_\Sigma \nabla_\Gamma \cdot \mathbf{q}\, d\Gamma,$$

this implies

$$V + \dot{u} + uVH + u\nabla_\Gamma \cdot \mathbf{T} = -\nabla_\Gamma \cdot \mathbf{q} + f, \tag{2.9}$$

which we can rewrite, by using (2.5), as

$$\partial_t u + \nabla_\Gamma \cdot (u\mathbf{T}) + uVH + V = -\nabla_\Gamma \cdot \mathbf{q} + f. \tag{2.10}$$

We now use (2.10) in (2.1) and obtain

$$\begin{aligned} \frac{d}{dt}E &= -\int_\Gamma (\nabla_\Gamma \cdot (u\mathbf{T}) + uVH + V + \nabla_\Gamma \cdot \mathbf{q} - f)\frac{\delta E}{\delta u}\, d\Gamma + \int_\Gamma \mathbf{v} \cdot \frac{\delta E}{\delta \Gamma}\, d\Gamma \\ &= -\int_\Gamma \nabla_\Gamma \cdot (\mathbf{q} + u\mathbf{T})\frac{\delta E}{\delta u}\, d\Gamma + \int_\Gamma f\frac{\delta E}{\delta u} + \int_\Gamma V\mathbf{n} \cdot \frac{\delta E}{\delta \Gamma} - (uHV + V)\frac{\delta E}{\delta u}\, d\Gamma, \end{aligned}$$

where we have used $\mathbf{T} \cdot \frac{\delta E}{\delta \Gamma} = 0$. We now define

$$\mathbf{q} + u\mathbf{T} = -\nu\nabla_\Gamma \frac{\delta E}{\delta u}, \tag{2.11}$$

$$f = -k\frac{\delta E}{\delta u}, \tag{2.12}$$

$$bV = -\left(\mathbf{n} \cdot \frac{\delta E}{\delta \Gamma} - (uH + 1)\frac{\delta E}{\delta u}\right), \tag{2.13}$$

with ν, k, and b positive coefficients. In a general setting these coefficients might depend on $\mathbf{n}$ and/or u. These definitions imply

$$\frac{d}{dt}E = -\int_\Gamma \frac{1}{\nu}(\mathbf{q} - u\mathbf{T})^2\, d\Gamma - \int_\Gamma \frac{1}{k} f^2\, d\Gamma - \int_\Gamma bV^2\, d\Gamma \leq 0 \tag{2.14}$$

and thus energy dissipation, and with it consistency with the second law of thermodynamics.

We now use (2.10) and (2.11)–(2.13) to obtain the desired evolution law,

$$\partial_t u + uVH + V = \nabla_\Gamma \cdot (\nu\nabla_\Gamma \mu) - k\mu, \tag{2.15}$$

$$\mu = \partial_u \gamma, \tag{2.16}$$

$$bV + H_{\gamma_1} + \gamma_2 H - \mu - uH\mu = 0, \tag{2.17}$$

which is exactly (1.7)–(1.9).

In the same way as the equations are derived from the energy $E[\Gamma, u] = \int_\Gamma \gamma \, d\Gamma$, with $\gamma = \gamma(\mathbf{n}, u)$, more general surface free energies can be used. One example would be a curvature regularized free energy, as originally proposed for the evolution of curves in [7] to deal with strong anisotropies with missing orientations, which lead to backward parabolic behavior of the evolution laws. An extension to surfaces is discussed in [10, 16]. Combined with the adatoms the energy reads

$$E[\Gamma, u] = \int_\Gamma \gamma + \epsilon^2 \frac{1}{2} H^2 \, d\Gamma,$$

with $\gamma = \gamma(\mathbf{n}, u)$. Here ϵ sets a new length scale over which corners and edges in the surface are smeared out. The corners and edges result from the nonconvexity of γ with respect to $\mathbf{n}$. Thus the energy can be seen as a geometric generalization of a Ginzburg–Landau-type energy, where the curvature H plays the role of the gradient term. The variational derivatives now read

$$\frac{\delta E}{\delta u} = \mu, \tag{2.18}$$

$$\frac{\delta E}{\delta \Gamma} = H_{\gamma_1} \mathbf{n} + \gamma_2 H \mathbf{n} - \epsilon^2 \left(\Delta_\Gamma H + H \left(\|S\|^2 - \frac{1}{2} H^2 \right) \right) \mathbf{n}, \tag{2.19}$$

with $S = \nabla_\Gamma \mathbf{n}$ being the shape operator and $\|S\| = \sqrt{\mathrm{trace}(SS^T)}$ its Frobenius norm. The resulting equations read

$$\partial_t u + uVH + V = \nabla_\Gamma \cdot (\nu \nabla_\Gamma \mu) - k\mu, \tag{2.20}$$

$$\mu = \partial_u \gamma, \tag{2.21}$$

$$bV + H_{\gamma_1} + \gamma_2 H - \epsilon^2 \left(\Delta_\Gamma H + H \left(\|S\|^2 - \frac{1}{2} H^2 \right) \right) = \mu + uH\mu. \tag{2.22}$$

If we consider (2.17) with given u and μ, we see that the equation can become backward parabolic if $H_{\gamma_1} + \gamma_2 H - uH\mu < 0$. This has already been discussed in the isotropic case in [3], where this situation might occur for large adatom densities and special choices for γ. In the anisotropic case this situation is much more likely, because H might be large in situations where H_γ is small. Thus even for weak anisotropies the discussed regularization might be necessary to deal with the resulting backward parabolic equation.

3. Numerical approach.

The system (2.15)–(2.17) with $H_{\gamma_1} = H$ and $k = 0$ has been numerically treated in a graph formulation by Burger [3] and in a phase field approach by Rätz and Voigt [17]. Here we will consider a different numerical approach and allow for anisotropies, including strong anisotropies. Thus we will solve the system (2.20)–(2.22). We will use an operator splitting ansatz and consider (2.20) as a diffusion equation on an evolving surface Γ, with Γ given, and (2.22) as the geometric equation which determines the normal velocity, with the interfacial quantities u and μ given.

3.1. Diffusion equations on evolving surface.

The systems (2.15)–(2.17) and (2.20)–(2.22) include the problem of solving a diffusion equation on an evolving surface. Such problems can be found in various applications, e.g., in materials science, biophysics, image processing, and computer graphics. Theoretical results for such problems are rare and also numerical methods are much less developed then for equations defined in $\mathbb{R}^d$. Only recently various numerical approaches are introduced. Dziuk and Elliott [8] introduced a direct approach by parametric finite elements, which allows one to solve a diffusion equation on an evolving surface. The evolution of the surface in their approach is given and for complicated movements severe problems in maintaining the regularity of the surface mesh have to be overcome. A different approach in which the surface is described implicitly, and thus problems with the regularity of surface meshes circumvented, was used by Xu and Zhao [21]. In Lowengrub, Xu, and Voigt [13] this approach is extended to solve a Cahn–Hilliard equation on an evolving surface. In both approaches the evolution of the surface is described through a level set function and results from the interaction with a surrounding flow field. Both implementations, however, are restricted to curves. A different approach in which the surface is described implicitly is used in Rätz and Voigt [17]. Here a phase field function is used to describe the evolving surface and the evolution is governed by a modified Allen–Cahn equation.

3.2. Level set approximation. We are going to use the level set method to solve diffusion equations on a stationary surface, introduced by Bertalmio et al. [2], and extend it to evolving surfaces. The interface is described implicitly through a level set function. Equation (2.20) is thereby approximated by the diffusion equation on the time-independent domain Ω, with $\Gamma(t) \subset \Omega$ for all $t \in (0, T)$,

$$\partial_t u - u\frac{\partial_t \psi}{\|\nabla\psi\|}\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|} - \frac{\partial_t\psi}{\|\nabla\psi\|} = \frac{1}{\|\nabla\psi\|}\nabla\cdot\big(\nu\|\nabla\psi\|P_{\nabla\psi}\nabla\mu\big) - k\mu, \tag{3.1}$$

where now u and μ denote the extended variable, defined on Ω, and $P_{\nabla\psi}$ is the projection operator on the tangential space of $\Gamma(t)$ defined through

$$P_{\nabla\psi} = id - \frac{\nabla\psi}{\|\nabla\psi\|}\otimes\frac{\nabla\psi}{\|\nabla\psi\|}.$$

The first terms on the left and right sides of (3.1) are shown to represent $\partial_t u = \nabla_\Gamma\cdot(\nu\nabla_\Gamma\mu)$; see [2]. In the second and third terms on the left side we use the level set equation $\partial_t\psi + V\|\nabla\psi\| = 0$ to obtain V and the definition of H in level set form. We obtain

$$V = -\frac{\partial_t\psi}{\|\nabla\psi\|},$$

$$H = \nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|}.$$

The second term on the left side can be rewritten as

$$u\partial_t\psi\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|} = \nabla\cdot\left(u\partial_t\psi\frac{\nabla\psi}{\|\nabla\psi\|}\right) - \partial_t\psi\nabla u\cdot\frac{\nabla\psi}{\|\nabla\psi\|} - u\nabla(\partial_t\psi)\cdot\frac{\nabla\psi}{\|\nabla\psi\|}.$$

For (2.22) we follow the level set representation for curvature regularized anisotropic mean curvature flow introduced by Burger et al. [5] and obtain

$$\begin{aligned}-\partial_t\psi + \frac{1}{b}\Bigg[\nabla\cdot\gamma_{1_z}\left(\frac{\nabla\psi}{\|\nabla\psi\|}\right) + \gamma_2(u)\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|} - \epsilon^2\left(\Delta_\Gamma H + H\left(\|S\|^2 - \frac{1}{2}H^2\right)\right)\\ - \mu - u\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|}\mu\Bigg]\|\nabla\psi\| = 0,\end{aligned} \tag{3.2}$$

with $\gamma_{1_z}\big(\frac{\nabla\psi}{\|\nabla\psi\|}\big) = D\gamma_1(\mathbf{n})$, where $D\gamma_1$ is the differential of the one-homogeneous function γ_1. H and S need to be replaced by their level set definitions, which will be done in the following weak formulation. The last term in (3.2) can be rewritten as

$$u\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|}\mu = \nabla\cdot\left(u\mu\frac{\nabla\psi}{\|\nabla\psi\|}\right) - \mu\nabla u\cdot\frac{\nabla\psi}{\|\nabla\psi\|} - u\nabla\mu\cdot\frac{\nabla\psi}{\|\nabla\psi\|},$$

and with (1.8) we have $\mu = \gamma_2'(u)$, which gives

$$\gamma_2(u)\nabla\cdot\frac{\nabla\psi}{\|\nabla\psi\|} = \nabla\cdot\left(\gamma_2(u)\frac{\nabla\psi}{\|\nabla\psi\|}\right) - \mu\nabla u\frac{\nabla\psi}{\|\nabla\psi\|}.$$

A weak form of (3.1) and (3.2) thus reads

$$\begin{aligned}
&\int_\Omega \|\nabla\psi\|\partial_t u\eta\,dx + \int_\Omega u\partial_t\psi\frac{\nabla\psi}{\|\nabla\psi\|}\cdot\nabla\eta\,dx + \int_\Omega \partial_t\psi\nabla u\cdot\frac{\nabla\psi}{\|\nabla\psi\|}\eta\,dx\\
&\qquad + \int_\Omega u\nabla(\partial_t\psi)\cdot\frac{\nabla\psi}{\|\nabla\psi\|}\eta\,dx - \int_\Omega \partial_t\psi\eta\,dx\\
&\quad = -\int_\Omega \nu\|\nabla\psi\|P_{\nabla\psi}\nabla\mu\cdot\nabla\eta\,dx - \int_\Omega k\|\nabla\psi\|\mu\eta\,dx,\\
&\int_\Omega \frac{b}{\|\nabla\psi\|}\partial_t\psi\xi\,dx + \int_\Omega \gamma_{1_z}\left(\frac{\nabla\psi}{\|\nabla\psi\|}\right)\cdot\nabla\xi\,dx + \int_\Omega \gamma_2(u)\frac{\nabla\psi}{\|\nabla\psi\|}\cdot\nabla\xi\,dx\\
&\qquad + \frac{\epsilon^2}{2}\int_\Omega \frac{\omega^2}{\|\nabla\psi\|^3}\nabla\psi\cdot\nabla\xi\,dx\\
&\qquad + \epsilon^2\int_\Omega \frac{P_{\nabla\psi}\nabla\omega}{\|\nabla\psi\|}\cdot\nabla\xi\,dx + \int_\Omega \mu\xi\,dx - \int_\Omega u\mu\frac{\nabla\psi}{\|\nabla\psi\|}\cdot\nabla\xi\,dx\\
&\qquad - \int_\Omega u\nabla\mu\cdot\frac{\nabla\psi}{\|\nabla\psi\|}\xi\,dx = 0,\\
&\int_\Omega \frac{\omega}{\|\nabla\psi\|}\varphi\,dx = \int_\Omega \frac{\nabla\psi}{\|\nabla\psi\|}\cdot\nabla\varphi\,dx,
\end{aligned}$$

with appropriate test functions η, ξ, and ϕ. The third equation results from setting $\omega = -\|\nabla\psi\|H$ in (3.2). To discretize in time we use an operator splitting approach, starting with the second and third equations, to obtain ψ^{n+1} by given ψ^n, u^n, and μ^n. We apply a convex splitting ansatz derived in [5] for the treatment of the nonlinear anisotropy term and obtain

$$\begin{aligned}
&\int_\Omega \frac{b}{\|\nabla\psi^n\|}\frac{\psi^{n+1}-\psi^n}{\tau^n}\xi\,dx + \int_\Omega \gamma_z\left(\frac{\nabla\psi^n}{\|\nabla\psi^n\|}\right)\cdot\nabla\xi\,dx + \int_\Omega \gamma_2(u^n)\frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\cdot\nabla\xi\,dx\\
&\qquad + \int_\Omega \frac{\lambda}{\|\nabla\psi^n\|}\gamma\left(\frac{\nabla\psi^n}{\|\nabla\psi^n\|}\right)(\nabla\psi^{n+1}-\nabla\psi^n)\cdot\nabla\xi\,dx\\
&\qquad + \frac{\epsilon^2}{2}\int_\Omega \frac{(\omega^n)^2}{\|\nabla\psi^n\|}\nabla\psi^{n+1}\cdot\nabla\xi\,dx + \epsilon^2\int_\Omega \frac{\omega^{n+1}}{\|\nabla\psi^n\|}\cdot\nabla\xi\,dx\\
&\qquad - \epsilon^2\int_\Omega \frac{\big(\mathrm{id}-P_{\nabla\psi^n}\big)\nabla\omega^n}{\|\nabla\psi^n\|}\cdot\nabla\xi\,dx + \int_\Omega \mu^n\xi\,dx - \int_\Omega u^n\mu^n\frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\cdot\nabla\xi\,dx
\end{aligned}$$

$$-\int_\Omega u^n\nabla\mu^n\cdot\frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\xi\,dx = 0, \tag{3.3}$$

$$\int_\Omega \frac{\omega^{n+1}}{\|\nabla\psi^n\|}\varphi\, dx = \int_\Omega \frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\cdot\nabla\varphi\, dx, \tag{3.4}$$

with λ an appropriate parameter. ψ^{n+1} is then used in the first equation to determine u^{n+1} by given u^n, ψ^n, and ψ^{n+1}:

$$\begin{aligned}
&\int_\Omega \|\nabla\psi^{n+1}\|\frac{u^{n+1}-u^n}{\tau^n}\, dx + \int_\Omega u^{n+1}\frac{\psi^{n+1}-\psi^n}{\tau^n}\frac{\nabla\psi^{n+1}}{\|\nabla\psi^{n+1}\|}\cdot\nabla\eta\, dx\\
&\quad + \int_\Omega \frac{\psi^{n+1}-\psi^n}{\tau^n}\nabla u^{n+1}\cdot\frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\eta\, dx + \int_\Omega u^{n+1}\frac{\nabla\big(\psi^{n+1}-\psi^n\big)}{\tau^n}\cdot\frac{\nabla\psi^{n+1}}{\|\nabla\psi^n\|}\eta\, dx\\
&\quad - \int_\Omega \frac{\psi^{n+1}-\psi^n}{\tau^n}\eta\, dx\\
&= -\int_\Omega \nu\|\nabla\psi^{n+1}\|P_{\nabla\psi}\nabla\mu^{n+1}\cdot\nabla\eta\, dx - \int_\Omega k\|\nabla\psi^{n+1}\|\mu^{n+1}\eta\, dx. \qquad (3.5)
\end{aligned}$$

For the calculation of u^{n+1} the adatom density u^n of the last timestep is needed at the new interface ψ^{n+1}. The interface moves in a normal direction. So to ensure that u^n has adequate values at the new interface, we extend the adatom density in direction normal to the interface after each timestep. The extension is done by an adaption of the redistancing algorithm introduced in [19] and described in detail in [4]. The approach uses the Hopf–Lax formula and offers an efficient alternative to fast marching or fast sweeping algorithms, which works also on unstructured grids.

The equations resulting from (3.3)–(3.5) are linear and are discretized in space by linear finite elements. To describe the linear system to be solved we use the following weighted mass, first order, and stiffness matrices:

$$M[f] := \left(\int_\Omega f\,\varphi_i\,\varphi_j\, dx\right)_{i,j}, \qquad F[v] := \left(\int_\Omega v\cdot\nabla\varphi_i\,\varphi_j\, dx\right)_{i,j},$$

$$L[f] := \left(\int_\Omega f\,\nabla\varphi_i\,\nabla\varphi_j\, dx\right)_{i,j}, \qquad L[A] := \left(\int_\Omega A\,\nabla\varphi_i\,\nabla\varphi_j\, dx\right)_{i,j},$$

with functions $f:\Omega\to\mathbb{R}$, $v:\Omega\to\mathbb{R}^d$ and $A:\Omega\to\mathbb{R}^{d\times d}$ and basis functions φ_i of the finite element space $\mathcal{V}^h$ defined through

$$\mathcal{V}^h = \{v\in C(\Omega)\mid v|_T \text{ is linear polynomial for } T\in\mathcal{T}\},$$

with $\mathcal{T}$ a decomposition of the polygonal domain Ω into triangles or tetrahedra and $\mathcal{V}_h\subset\mathcal{V}=H^1(\Omega)$. In the case of periodic boundary conditions on part of the boundary $\Gamma_{per}\subset\partial\Omega$, the corresponding periodic subspaces of $\mathcal{V}$ and $\mathcal{V}_h$ are used. Moreover we assume natural boundary conditions for all variables on the boundary $\partial\Omega\backslash\Gamma_{per}$. We make use of a norm regularization $\|\nabla\psi^n\|_\delta = (\|\nabla\psi^n\|^2+\delta^2)^{1/2}$, δ about the grid size, for small norm values $\|\nabla\psi^n\|$. With the matrices and right-hand side vectors

$$\begin{aligned}
M_1 &:= M[\|\nabla\psi^n\|_\delta^{-1}], & M_2 &:= M[1],\\
M_3 &:= M[b\,\|\nabla\psi^n\|_\delta^{-1}], & M_4 &:= M[k\,\|\nabla\psi^n\|_\delta],\\
M_5 &:= M[\|\nabla\psi^{n+1}\|_\delta], & M_6 &:= M[\|\nabla\psi^{n+1}\|_\delta^{-1}\,\nabla\psi^{n+1}\cdot\nabla\big(\psi^{n+1}-\psi^n\big)],\\
F_1 &:= F[u^n\,\nabla\mu^n\,\|\nabla\psi^n\|_\delta^{-1}], & F_2 &:= F[\big(\psi^{n+1}-\psi^n\big)\,\|\nabla\psi^{n+1}\|_\delta^{-1}\,\nabla\psi^{n+1}],\\
L_1 &:= L[(W^n)^2\,\|\nabla\psi^n\|_\delta^{-3}], & L_3 &:= L[\|\nabla\psi^n\|_\delta^{-1}],\\
L_4 &:= L[(id-P_{\nabla\psi^n})\,\|\nabla\psi^n\|_\delta^{-1}], & L_5 &:= L\left[\gamma\left(\frac{\nabla\psi^n}{\|\nabla\psi^n\|_\delta}\right)\,\|\nabla\psi^n\|_\delta^{-1}\right],\\
L_6 &:= L[\nu\,P_{\nabla\psi^n}\,\|\nabla\psi^n\|_\delta], & L_7 &:= L[u^n\,\mu^n\,\|\nabla\psi^n\|_\delta^{-1}],\\
L_8 &:= L[(u^n)^2\,\|\nabla\psi^n\|_\delta^{-1}], & G &:= \left(\int_\Omega \gamma_z\left(\frac{\nabla\psi^n}{\|\nabla\psi^n\|_\delta}\right)\cdot\nabla\varphi_j\,dx\right)_j,
\end{aligned}$$

and the linear expansions

$$\psi^{n+1}=\sum_{i=1}^{L}\Psi_i^{n+1}\varphi_i,\qquad \omega^{n+1}=\sum_{i=1}^{L}W_i^{n+1}\varphi_i,$$
$$\mu^{n+1}=\sum_{i=1}^{L}\Upsilon_i^{n+1}\varphi_i,\qquad u^{n+1}=\sum_{i=1}^{L}U_i^{n+1}\varphi_i,$$

with L the dimension of the finite element space $\mathcal{V}^h$, the linear system for (3.3) and (3.4) then reads

$$\begin{pmatrix} \begin{array}{c} M_3+\tau\frac{\epsilon^2}{2}L_1+\tau\lambda L_5\\ -\tau L_7+\tau\frac{\alpha}{2}L_8-\tau F_1\end{array} & \tau\epsilon^2 L_3\\ -L_3 & M_1\end{pmatrix}\begin{pmatrix}\Psi^{n+1}\\ W^{n+1}\end{pmatrix}=\begin{pmatrix}\begin{array}{c} -\tau M_2\Upsilon^n+M_3\Psi^n+\tau\epsilon^2L_4W^n\\ +\tau\lambda L_5\Psi^n-\tau G\end{array}\\ 0\end{pmatrix}.$$

We use a Schur complement approach to solve the system for the unknown ψ^{n+1}, which gives

$$\begin{aligned}
&\left(M_3+\tau\frac{\epsilon^2}{2}L_1+\tau\lambda L_5-\tau L_7+\tau\frac{\alpha}{2}L_8-\tau F_1+\tau\epsilon^2L_3M_1^{-1}L_3\right)\Psi^{n+1}\\
&\qquad=-\tau M_2\Upsilon^n+M_3\Psi^n+\tau\epsilon^2L_4W^n+\tau\lambda L_5\Psi^n-\tau G,
\end{aligned}$$

where W^n is calculated via $W^n=M_1^{-1}L_3\psi^n$ and the inverse M_1^{-1} is obtained with mass lumping. The Schur complement system is solved by a GMRES solver, as the system matrix might not be positive definite for the timesteps used. Within the special case of a free energy γ such that $\mu=\alpha u$, with α a positive parameter, the system for (3.5) reads

$$\big[M_5+\big(F_2+F_2^T\big)+M_6+\tau\alpha L_6+\tau\alpha M_4\big]\,U^{n+1}=M_5U^k+M_2\big(\Psi^{k+1}-\Psi^k\big)$$

and is solved by a GMRES solver.

4. Simulation results. The derived numerical scheme is implemented in the adaptive finite element toolbox AMDiS [20]. The toolbox provides a framework for the efficient solution of systems of partial differential equations by adaptive finite

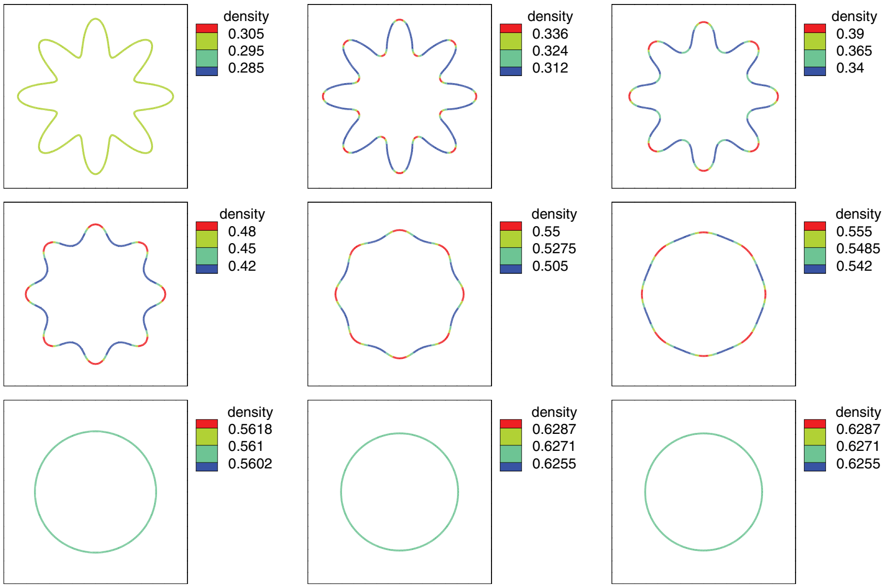


FIG. 4.1. *Isotropic evolution of interface and adatom density. Simulation parameters:* $[0,4] \times [0,4]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.3$*. From top left to bottom right:* $t = 0.0$, 0.002, 0.01, 0.03, 0.06, 0.09, 0.2, 2.5, 3.0.

elements. In the following examples linear finite elements are used on unstructured simplicial meshes, and the refinement is based on bisection. The criteria for refinement/coarsening is heuristic and based on the distance to the zero level set. This reduces the computational overhead by allowing for coarse meshes away from the zero level set. The computational cost is therefore comparable to a narrow band approach or a phase-field approximation with adaptive refinement at the diffuse interface.

The examples are chosen such that the influence of the adatom density is highlighted. For that purpose we compare the results with the corresponding model (1.5) in which the adatom density is neglected. The numerical approach used to solve (1.5) is described in [18]. In all simulations we use either the isotropic function $\gamma_1(\mathbf{n}) = 1$ or the anisotropic function

$$\gamma_1(\mathbf{n}) = 1 + a \sum_{k=1}^{d} n_k^4, \quad d = 2, 3,$$

with n_k denoting the kth spatial component of the normal. If not stated otherwise, we choose $a = 2.0$ (i.e., γ_1 is nonconvex), $\alpha = 1.0$ (i.e., $\mu = u$), the regularization parameter $\epsilon = 0.1$, the evaporation modulus $k = 0$, the kinetic coefficient $b = 1.0$, and the surface mobility $\nu = 1.0$. The relation between grid size h at the interface and timestep Δt is $\Delta t \leq h^4/\epsilon^2$. As ϵ has to be resolved by h we obtain with $h \sim \epsilon$ a timestep restriction of approximately $\Delta t \leq h^2$, which justifies the use of a semi-implicit discretization, as the restriction for an explicit strategy would be $\Delta t \leq h^6$ for the underlying sixth order problem.

Figure 4.1 shows the evolution of a perturbed circle with initial constant adatom concentration to the Wulff shape under isotropic conditions. The results are in agreement with the phase-field simulations for the same problem considered in [17]. The

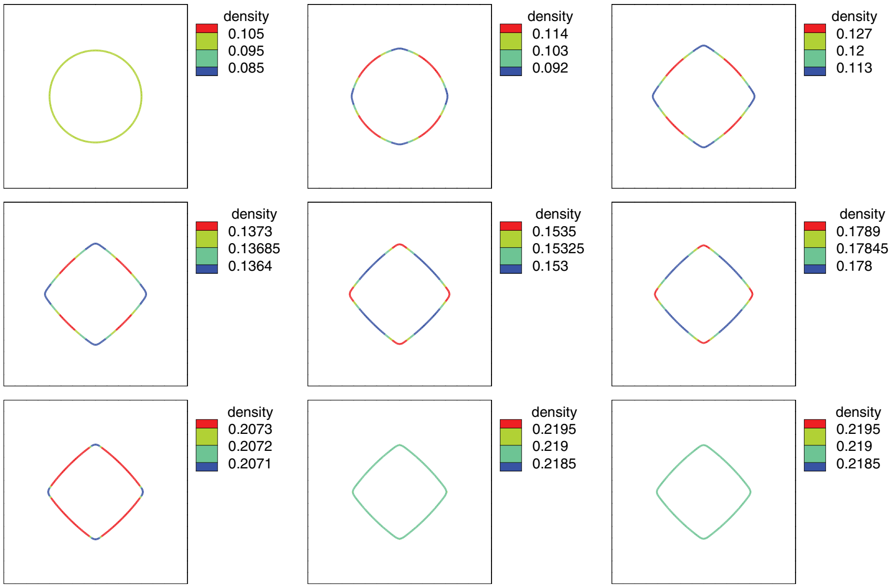


FIG. 4.2. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[0,4] \times [0,4]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.1$*,* $\alpha = 10.0$*. From top left to bottom right:* $t = 0.0$, 0.005, 0.02, 0.04, 0.06, 0.1, 0.2, 0.5, 1.0.

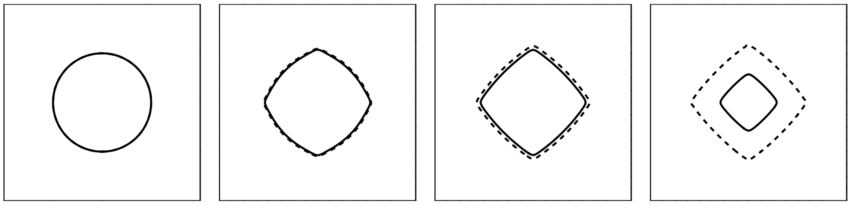


FIG. 4.3. *Anisotropic evolution: comparison of adatom model (solid) and kinetic model (dashed). Simulation parameters:* $[0,4] \times [0,4]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$ *(adatom model), and* $\Delta t = 5 \times 10^{-6}$ *(kinetic model). In adatom model:* $u_0 = 0.1$*. From left to right:* $t = 0.0$, 0.01, 0.03, 0.2.

perturbations smooth out and influence the adatom concentration, which adjusts to the local curvature and velocity. After a circle is obtained the adatom concentration becomes constant and converges to an equilibrium value, which is determined through the initial conditions.

Figure 4.2 shows the evolution of a circle with a constant initial adatom concentration to its Wulff shape under anisotropic conditions. Again the adatom concentration adjusts during the evolution to the local curvature and velocity, but if the Wulff shape is reached the adatom concentration becomes constant and converges towards its equilibrium value, which again depends on the initial conditions. Thus in equilibrium the adatom concentration is constant, independent on the local curvature.

If we compare this evolution with the kinetic surface diffusion model in which adatoms are neglected, we obtain similar results on the dynamics and the equilibrium shape; see Figure 4.3. The only difference is the size of the final shape, as in the kinetic surface diffusion model conservation in mass results in conservation in area, which is

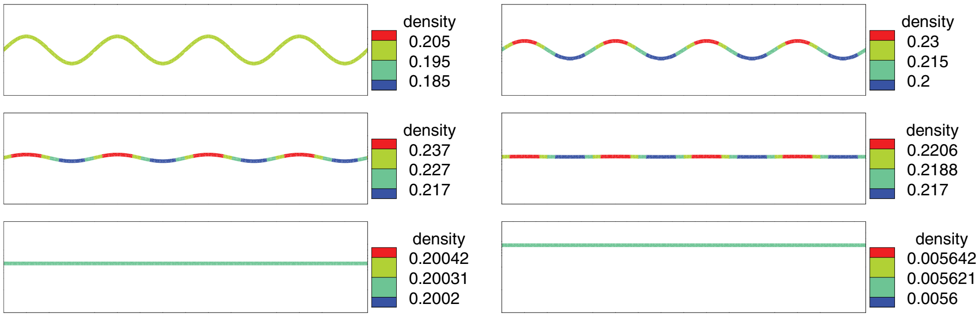


FIG. 4.4. *Isotropic evolution of interface and adatom density. Simulation parameters:* $[0,4] \times [-0.5,0.5]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0$, 0.01, 0.03, 0.09, 0.2, 7.0.

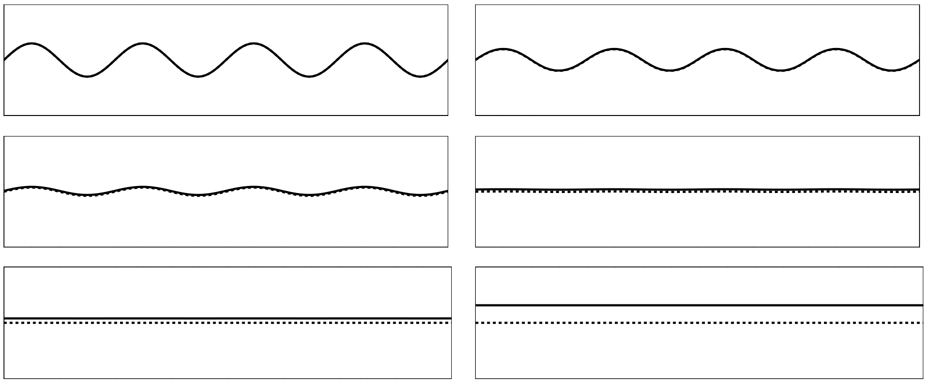


FIG. 4.5. *Isotropic evolution: comparison of adatom model (solid) and kinetic model (dashed). Simulation parameters:* $[0,4] \times [-0.5,0.5]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$ *(adatom model), and* $\Delta t = 5 \times 10^{-6}$ *(kinetic model). In adatom model:* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0$, 0.01, 0.03, 0.09, 0.2, 1.3.

not the case in the adatom model. Here the area can be reduced by increasing the adatom density. From a numerical point of view an additional difference is observed. The timesteps in the adatom model can be chosen larger than for the kinetic surface diffusion model. This results from the diffusion character of the adatom model and has already been speculated in [9].

The following examples are devoted to curves, which are not closed. Figure 4.4 shows the evolution of a perturbed straight line with a constant initial adatom concentration under isotropic conditions. In agreement with the results obtained by a graph formulation [3] and a phase-field formulation [17] the perturbations smooth out and the adatom concentration adjusts to the local curvature and velocity. After a straight line is formed the adatom concentration converges to zero.

In Figure 4.5 the evolution is compared with the kinetic surface diffusion model in which adatoms are neglected. Again the results agree, with the only difference being that the height in the adatom model is higher, which results from the additional mass from the adatoms, which become incorporated during the evolution. Numerically we again can use larger timesteps in the adatom model.

Figure 4.6 shows the evolution of a perturbed straight line with a constant initial

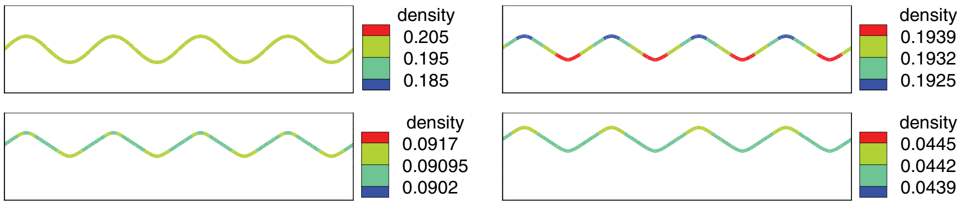


FIG. 4.6. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[0,4]\times[-1.0,1.0]$ *grid (shown here:* $[0,4]\times[-0.5,0.5]$*), grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0$, 0.001, 1.0, 2.0.

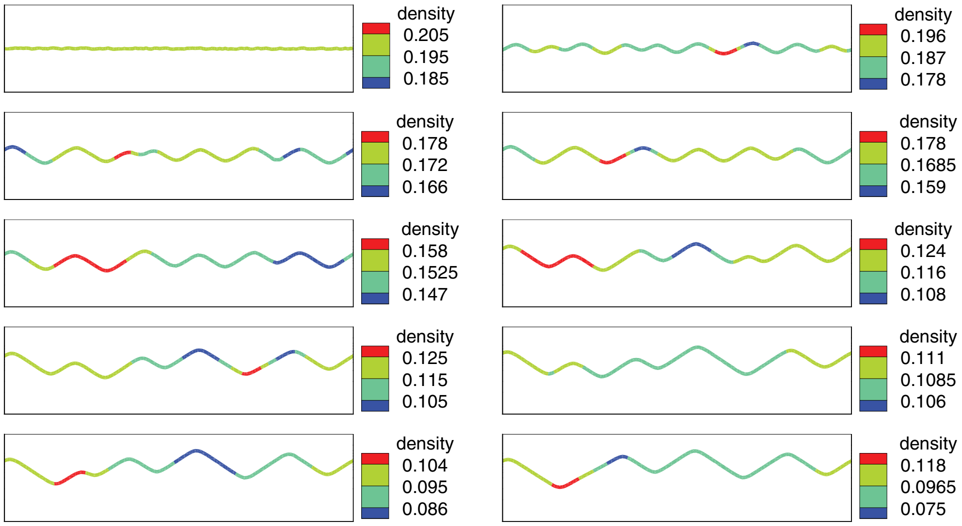


FIG. 4.7. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[0,4]\times[-1.0,1.0]$ *grid (shown here:* $[0,4]\times[-0.5,0.5]$*), grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0$, 0.02, 0.04, 0.06, 0.07, 0.2, 0.76, 0.8, 1.2, 2.6.

adatom concentration under anisotropic conditions. We observe the facet formation and the adjustment of the adatom concentration to the local curvature and velocity. After the facets are formed the adatom concentration is reduced. The structure, however, is not stable, due to the corner energy resulting from the regularization. Thus coarsening is expected, which will reduce the number of corners.

To further study the coarsening, we now start from an initially unstable orientation with an initially constant adatom concentration. Figures 4.7 and 4.8 show the spinodal decomposition into allowed orientations and the subsequent coarsening of the structure. The influence of the coarsening event on the adatom concentration can clearly be observed. During coarsening the interface moves fast. High velocities in the coarsening areas result in a smaller adatom concentration where the interface moves in a positive normal direction and a higher adatom concentration where the interface moves in a negative normal direction.

Again we compare this evolution with the kinetic surface diffusion model. Figure 4.9 shows the qualitative agreement. Numerically we again can use larger timesteps in the adatom model.

Finally we show results in three dimensions. We start with a cube and a constant

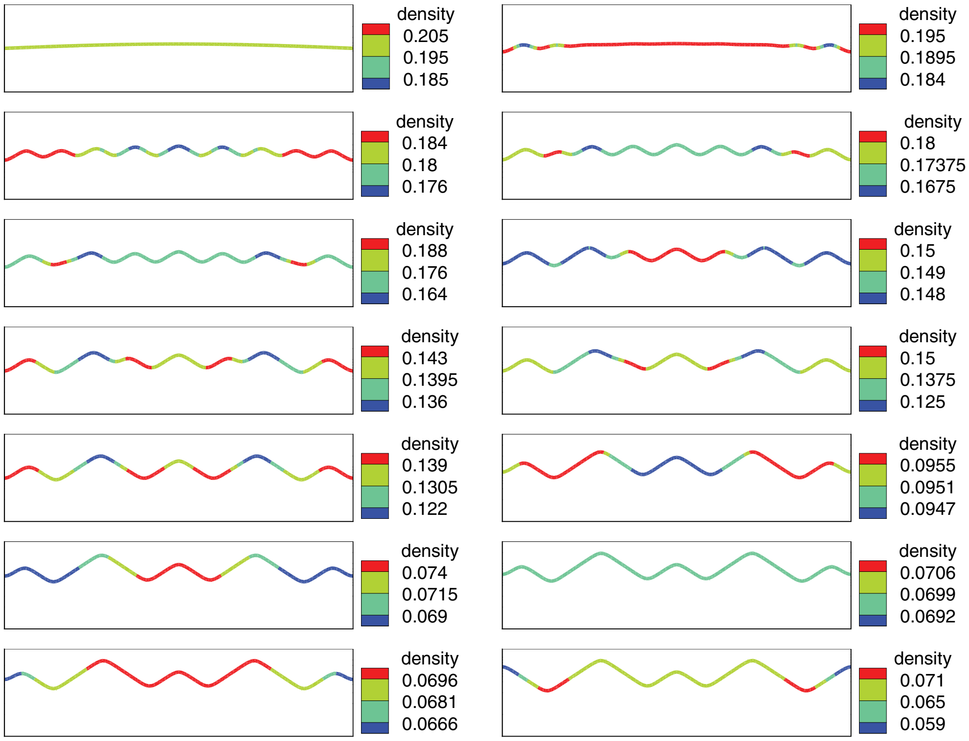


FIG. 4.8. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[0, 4] \times [-1.0, 1.0]$ *grid (shown here:* $[0, 4] \times [-0.5, 0.5]$*), grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0, 0.008, 0.02, 0.07, 0.08, 0.3, 0.4, 0.42, 0.43, 0.9, 1.4, 2.4, 2.5, 2.6$.

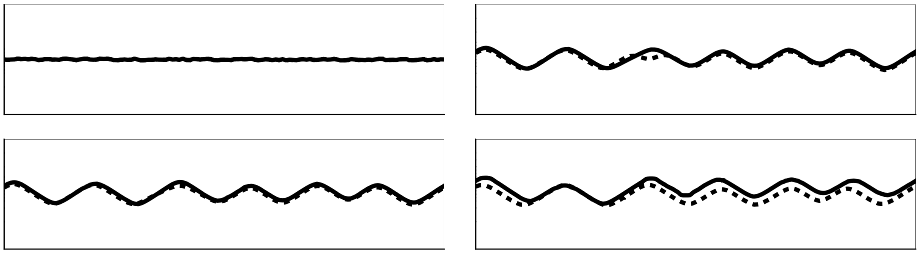


FIG. 4.9. *Anisotropic evolution: comparison of adatom model (solid) and kinetic model (dashed). Simulation parameters:* $[0, 4] \times [-0.5, 0.5]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$ *(adatom model), and* $\Delta t = 5 \times 10^{-6}$ *(kinetic model). In adatom model:* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0, 0.07, 0.08, 0.4$.

adatom density, which relaxes to sphere. The adatom density adjusts to local curvature and velocity during evolution and saturates at a constant value; see Figure 4.10. Figure 4.11 shows the evolution of a randomly perturbed initial surface within an unstable orientation and a constant initial adatom concentration. Again the facet formation and the subsequent coarsening of the surface morphology can be observed. The adatom concentration adjusts to the local curvature and velocity. If we start with a periodic structure and a constant initial adatom concentration, we observe a sym-

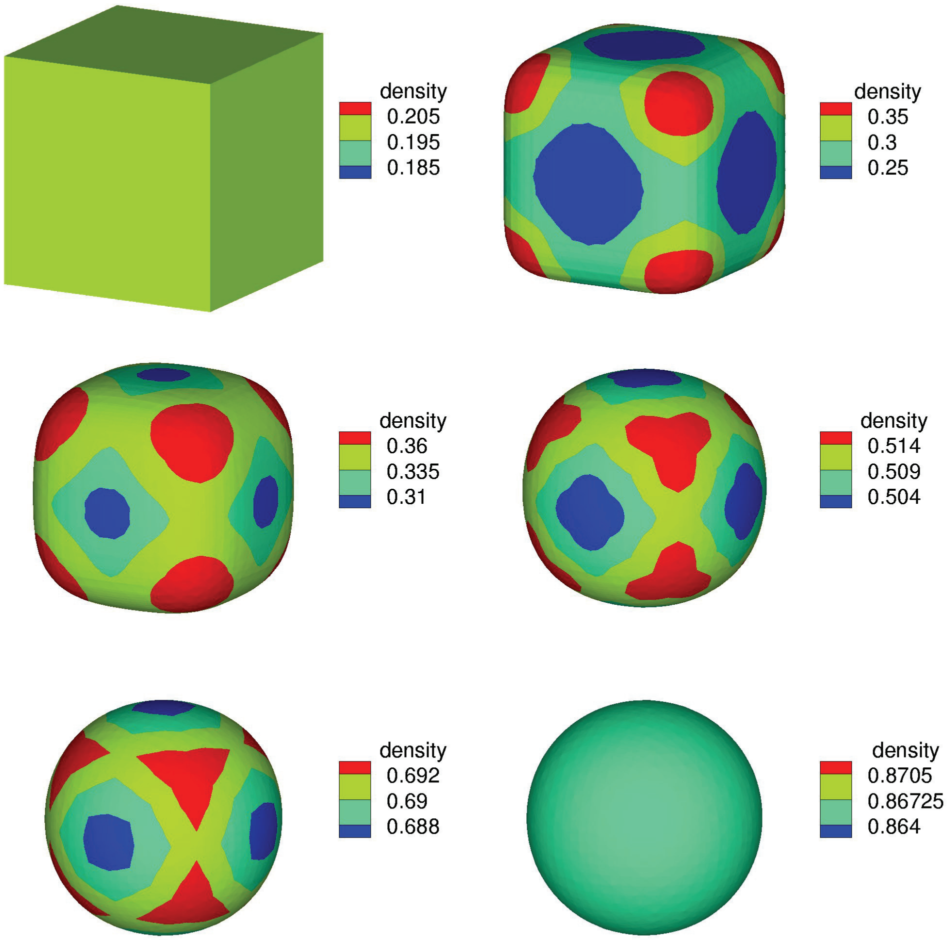


FIG. 4.10. *Isotropic evolution of interface and adatom density. Simulation parameters:* $[0.0, 4.0] \times [0.0, 4.0] \times [0.0, 4.0]$ *grid, grid size* $h = 0.0625$*, timestep* $\Delta t = 10^{-5}$*,* $\epsilon = 0.0$*,* $u_0 = 0.2$*. From top left to bottom right:* $t = 0.0$, 0.02, 0.05, 0.1, 0.15, 0.2.

metric coarsening in which four mounds collapse to form one mound; see Figure 4.12. This symmetry is probably a result of the 4-fold symmetry of the anisotropy function. A further example in Figure 4.10 shows the evolution of a cube with initial constant adatom concentration under isotropic conditions.

A more detailed study on the coarsening with more realistic physical parameters also under growth will be done elsewhere. Of interest is the influence of the adatom density on coarsening laws, as derived by Haußer and Voigt [11, 12]. The simulations in this paper are depicted to show the applicability of the algorithm. It has been shown that the introduced level set approach can be used to efficiently solve evolution equations on evolving surfaces. The approach is not restricted to diffusion problems on evolving surfaces but can be generalized to other equations as well.

Acknowledgment. The authors thank Martin Burger for helpful discussions.

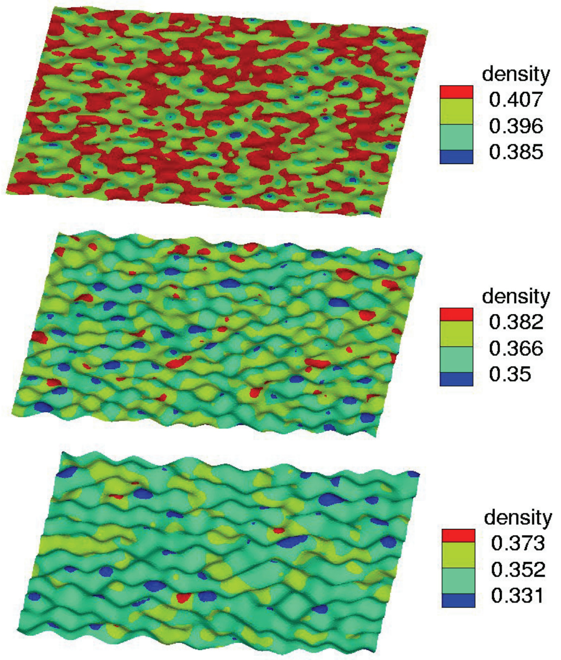


FIG. 4.11. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[-2.0, 2.0] \times [-2.0, 2.0] \times [-0.5, 0.5]$ *grid, grid size* $h = 0.03125$*, timestep* $\Delta t = 10^{-5}$*,* $\epsilon = 0.07$*,* $u_0 = 0.4$*. From top to bottom:* $t = 0.001, 0.005, 0.014$.

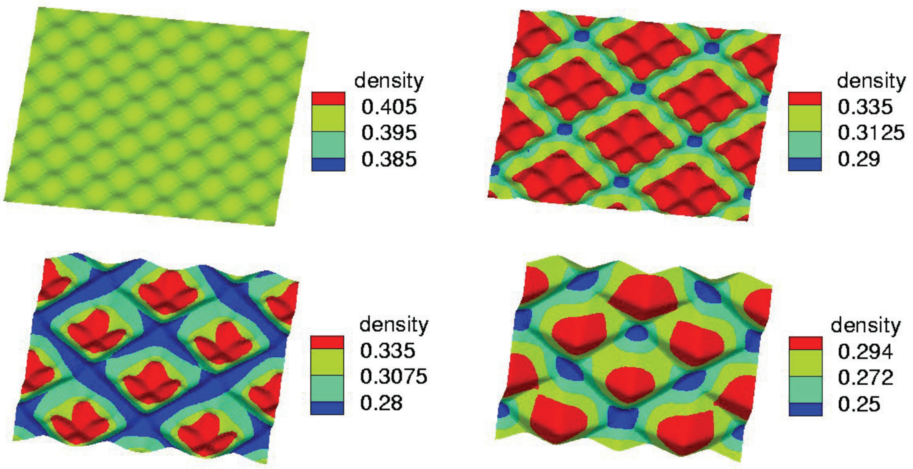


FIG. 4.12. *Anisotropic evolution of interface and adatom density. Simulation parameters:* $[-2.0, 2.0] \times [-2.0, 2.0] \times [-0.5, 0.5]$ *grid, grid size* $h = 0.05$*, timestep* $\Delta t = 10^{-4}$*,* $\epsilon = 0.07$*,* $u_0 = 0.4$*. From top left to bottom right:* $t = 0.0, 0.05, 0.1, 0.15$.

REFERENCES

[1] S. ANGENENT AND M. E. GURTIN, *Multiphase thermomechanics with interfacial structure,* 2: *Evolution of an isothermal surface*, Arch. Ration. Mech. Anal., 108 (1989), pp. 323–391.

[2] M. BERTALMIO, L. T. CHENG, S. OSHER, AND G. SAPIRO, *Variational problems and partial dif-*

ferential equations on implicit surfaces: The framework and examples in image processing and pattern formation, J. Comput. Phys., 174 (2001), pp. 759–780.

[3] M. Burger, *Surface diffusion including adatoms*, Commun. Math. Sci., 4 (2006), pp. 1–51.

[4] M. Burger, C. Stöcker, and A. Voigt, *Finite element based level set methods for higher order flows*, J. Sci. Comput., to appear.

[5] M. Burger, F. Haußer, C. Stöcker, and A. Voigt, *A level set approach to anisotropic flows with curvature regularization*, J. Comput. Phys., 225 (2007), pp. 183–205.

[6] J. W. Cahn and J. E. Taylor, *Surface motion by surface diffusion*, Acta Metall. Mater., 42 (1994), pp. 1045–1063.

[7] A. Di Carlo, M. E. Gurtin, and P. Podio-Guidugli, *A regularized equation for anisotropic motion-by-curvature*, SIAM J. Appl. Math., 52 (1992), pp. 1111–1119.

[8] G. Dziuk and C. Elliott, *Finite elements on evolving surfaces*, IMA J. Numer. Anal., 27 (2007), pp. 262–292.

[9] E. Fried and M. E. Gurtin, *A unified treatment of evolving interfaces accounting for small deformations and atomic transport with emphasis on grain-boundaries and epitaxy*, Adv. Appl. Mech., 40 (2004), pp. 1–177.

[10] M. E. Gurtin and M. E. Jabbour, *Interface evolution in three dimensions with curvature-dependent energy and surface diffusion: Interface-controlled evolution, phase transitions, epitaxial growth of elastic films*, Arch. Ration. Mech. Anal., 163 (2002), pp. 171–208.

[11] F. Haußer and A. Voigt, *Facet formation and coarsening modeled by a geometric evolution law for epitaxial growth*, J. Crystal Growth, 275 (2005), pp. e47–e51.

[12] F. Haußer and A. Voigt, *A geometric Ginzburg-Landau theory for faceted crystals in* 1*d: From coarsening to chaos through a driving force*, Phys. Rev. E, submitted.

[13] J. Lowengrub, J. J. Xu, and A. Voigt, *Surface phase separation and flow in a simple model of multicomponent drops and vesicles*, Fluid Dyn. Mater. Proc., 3 (2007), pp. 1–19.

[14] W. W. Mullins, *Two-dimensional motion of idealized grain boundaries*, J. Appl. Phys., 27 (1956), pp. 900–904.

[15] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.

[16] A. Rätz and A. Voigt, *Higher order regularization of anisotropic geometric evolution equations in three dimensions*, J. Comput. Theor. Nanosci., 3 (2006), pp. 560–564.

[17] A. Rätz and A. Voigt, *A diffuse-interface approximation for surface diffusion including adatoms*, Nonlinearity, 20 (2007), pp. 177–192.

[18] C. Stöcker and A. Voigt, *The effect of kinetics in the surface evolution of thin crystalline films*, J. Crystal Growth, 303 (2007), pp. 90–94.

[19] C. Stöcker, S. Vey, and A. Voigt, *Adaptive multidimensional simulations: Composite finite elements and signed distance functions*, WSEAS Trans. Circuits Syst., 4 (2005), pp. 111–116.

[20] S. Vey and A. Voigt, *AMDiS—adaptive multidimensional simulations*, Comput. Vis. Sci., 10 (2007), pp. 57–67.

[21] J. J. Xu and H. K. Zhao, *An Eulerian formulation for solving partial differential equations along a moving interface*, J. Sci. Comput., 19 (2003), pp. 573–594.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 81–110

INVERSE SOURCE PROBLEM IN NONHOMOGENEOUS BACKGROUND MEDIA. PART II: VECTOR FORMULATION AND ANTENNA SUBSTRATE PERFORMANCE CHARACTERIZATION[*]

EDWIN A. MARENGO[†], MOHAMED R. KHODJA[†], AND ABDELKADER BOUCHERIF[‡]

Abstract. This paper solves analytically and illustrates numerically the full-vector, electromagnetic inverse source problem of synthesizing an unknown source embedded in a given substrate medium of volume V and radiating a prescribed exterior field. The derived formulation and results generalize previous work on the scalar version of the problem, especially the recent Part I of this paper [A. J. Devaney, E. A. Marengo, and M. Li, *SIAM J. Appl. Math.*, 67 (2007), pp. 1353–1378]. Emphasis is put on substrates having constant constitutive properties within the source volume V, which, for formal tractability, is taken to be of spherical shape. The adopted approach is one of constrained optimization which also relies on spherical wavefunction theory. We find that the observed peaks in the spectrum of the singular values are primarily due to the phenomenon of Mie resonance. Therefore, for a given antenna radiating at a prescribed frequency, the set of solutions to the Mie resonance conditions corresponds to a set of constitutive parameters that maximize the radiated electromagnetic fields. The derived theory and associated implications for antenna substrates are illustrated numerically.

Key words. inverse source problem, antenna substrate, antenna performance, antenna limits, minimum energy, reactive power, antenna synthesis

AMS subject classification. 78-02

DOI. 10.1137/070689875

1. Introduction.

In this paper we investigate the full-vector, electromagnetic inverse source problem of reconstructing an unknown source (antenna) that is embedded, within a spherical region $V \equiv \left\{\mathbf{r} \in \mathbb{R}^3 : r \equiv |\mathbf{r}| \leq a\right\}$, in a given material or metamaterial substrate, and that radiates a given exterior field outside V. The derived formulation and results of this *inverse source problem in substrate media* generalize, within the full-vector formulation, previous work on the inverse source problem in free space (cf. [32] and the references therein), as well as previous work on the scalar version of the problem for nonhomogeneous backgrounds [13, 43], particularly Part I of this paper [12], coauthored by one of the authors of the present paper (Marengo). The formulation is based on constrained optimization. Two solution constraints are emphasized in the paper, in particular, the minimizing of the source L^2-norm or functional energy characterizing the "current level," with and without tuning to resonance, the former case corresponding to zero source reactive power. The ability of an antenna to radiate a prescribed power with reduced current levels as characterized by this norm is an indication of efficiency which has been adopted as a constraint in the antenna synthesis problem [5, 9]. Fundamental radiation limits related to the realizability of given fields or radiation performance with given source resources (antenna size, current level as measured by the source energy, reactive power, and so on)

[*] Received by the editors April 28, 2007; accepted for publication (in revised form) May 22, 2008; published electronically August 27, 2008. This work was supported by the United States Air Force Office of Scientific Research under grant FA9550-06-01-0013.

http://www.siam.org/journals/siap/69-1/68987.html

[†] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 (emarengo@ece.neu.edu, khodja.m@neu.edu).

[‡] Division of Applied Mathematics, Brown University, Providence, RI 02912, and King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia (Abdelkader_Boucherif@brown.edu, aboucher@kfupm.edu.sa).

or, alternatively, of the minimal resources needed for a given performance, are also elucidated as a by-product of the derived inverse source theory.

Motivation for this research is provided by the possibility of embedding an antenna in a substrate of a given size, where the original antenna plus the substrate are treated as the total antenna, so as to generate a given field or performance level which could not be achieved under the same physical constraints by another antenna in free space (i.e., without the substrate medium). This possibility has attracted interest from time to time in the antenna community; of particular interest have been a variety of antenna-embedding materials, including plasmas [37], nonmagnetic dielectrics [24, 25, 27, 26, 22, 2], magneto-dielectrics [18, 10, 36, 29], and, more recently, double-negative and single-negative metamaterials, which are receiving much attention as antenna performance-enhancing substrates by a number of groups [41, 40, 16, 3, 48, 42, 17, 28, 47]. The envisaged property is miniaturization of antennas by controlling electric size, via larger wavenumber, but other effects are involved, particularly when metamaterials are used. (A review of the pertinent state-of-the-art can be found in [15].)

For instance, it is well known [31, 34] that in the free-space case the source energy increases exponentially, for a given radiation pattern, with decreasing k_0a, where k_0 is the free-space wavenumber of the field. This increase occurs below a critical point determined by the fine detail that is desired in the radiation pattern, specifically, the antenna directivity. The question then is whether the critical source size in question can be made smaller by embedding the source in a properly selected substrate which becomes integral to the antenna. For small antennas (whose dimensions are smaller than about 1/3 of the wavelength [45, 46]) one is particularly interested in achieving radiation of an elemental dipolar mode, using minimal resources. Can antenna substrates help toward this goal? Alternatively, in certain applications using larger, resonant antennas whose dimensions are comparable to or larger than the wavelength, one can dispose of some “extra space” to accommodate a substrate, and the question is, Does antenna embedding yield enhancement of antenna directivity? Which values of the constitutive parameters give better performance?

We address these and related questions aided by the formalism of the inverse source problem, paying particular attention to lossless piecewise-constant radially symmetric backgrounds having electric permittivity ϵ_s and permeability μ_s. In particular, the total permittivity distribution is of the form

$$\epsilon(\mathbf{r}) = \epsilon_s\Theta(a-r) + \epsilon_0\Theta(r-a), \tag{1.1}$$

where Θ denotes Heaviside's unit step function ($\Theta(x) = 1$ for $x \geqslant 1$; otherwise $\Theta(x) = 0$), and the total permeability distribution is of the form

$$\mu(\mathbf{r}) = \mu_s\Theta(a-r) + \mu_0\Theta(r-a). \tag{1.2}$$

All the results are derived for time-harmonic fields, and thus the values of the constitutive parameters, which generally vary with frequency, are considered in this work for a given central frequency only.

Our results reveal the performance improvements due to antenna-embedding substrates from a fundamental inverse antenna theory point of view which is different from and complementary to efforts by other groups in this fruitful area. The observed peaks in the spectrum of the singular values of the source-to-exterior field mapping are primarily due to the phenomenon of Mie resonance, and maximum enhancement conditions are effectively summarized by the Mie resonance conditions (4.2) and (4.3).

Therefore, for a given antenna radiating at a prescribed frequency, the set of solutions to (4.2) and (4.3) corresponds to a set of constitutive parameters that maximize the radiated electromagnetic fields. As their amplitudes increase these radiated fields draw energy from the embedding medium. But because this medium is of finite extent the energy extraction process saturates, ultimately, and as a result of this saturation the fields fall short of effectively "blowing up."

2. The forward problem.

2.1. Electromagnetic generalities. Our starting point is provided by the frequency-domain Maxwell equations for a generally lossless, nonhomogeneous medium, in particular [7, 11],

$$\nabla \times \mathbf{E}(\mathbf{r}) = i\omega\mu(\mathbf{r})\mathbf{H}(\mathbf{r}), \tag{2.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = \mathbf{J}(\mathbf{r}) - i\omega\epsilon(\mathbf{r})\mathbf{E}(\mathbf{r}), \tag{2.2}$$

where $\mathbf{J}(\mathbf{r})$ represents an impressed current density (i.e., the source) confined within the spherical volume V, and $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ are, respectively, the electric and magnetic fields it generates. (These fields are subject to the radiation condition [38].) Substituting $\mathbf{H}(\mathbf{r})$ from (2.1) into (2.2) yields the vector wave equation

$$\nabla \times \left(\frac{\nabla \times \mathbf{E}(\mathbf{r})}{\mu(\mathbf{r})} \right) - \omega^2\epsilon(\mathbf{r})\mathbf{E}(\mathbf{r}) = i\omega\mathbf{J}(\mathbf{r}). \tag{2.3}$$

The partial differential operator in (2.3) admits an outgoing-wave dyadic Green's function $\bar{\mathbf{G}}(\mathbf{r}, \mathbf{r}')$ which, along with the radiation condition, obeys

$$\nabla \times \left(\frac{\nabla \times \bar{\mathbf{G}}(\mathbf{r}, \mathbf{r}')}{\mu(\mathbf{r})} \right) - \omega^2\epsilon(\mathbf{r})\bar{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = i\omega\delta(\mathbf{r} - \mathbf{r}')\bar{\mathbf{I}}, \tag{2.4}$$

where $\bar{\mathbf{I}}$ denotes the identity dyadic and δ the Dirac delta.

For future convenience we define the weighted inner product

$$(\mathbf{f}, \mathbf{f}') = \int d\mathbf{r} M(\mathbf{r}) \mathbf{f}^*(\mathbf{r}) \cdot \mathbf{f}'(\mathbf{r}), \tag{2.5}$$

where $\mathbf{f}$ and $\mathbf{f}'$ are any two functions of position and the asterisk $*$ denotes the complex conjugate; $M(\mathbf{r})$ is a characteristic (indicator or masking) function defined as

$$M(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in V, \\ 0, & \mathbf{r} \notin V. \end{cases} \tag{2.6}$$

Using this inner product, we express the source energy $\mathcal{E}$ as

$$\mathcal{E} \equiv (\mathbf{J}, \mathbf{J}), \tag{2.7}$$

and the complex interaction power $\mathcal{P}$ (cf. [7]) as

$$\mathcal{P} = -\frac{1}{2}(\mathbf{J}, \widetilde{\mathbf{G}}\mathbf{J}), \tag{2.8}$$

where we have introduced the linear mapping $\widetilde{\mathbf{G}}$ defined by

$$[\widetilde{\mathbf{G}}\mathbf{J}](\mathbf{r}) \equiv \int d\mathbf{r}' \bar{\mathbf{G}}(\mathbf{r}, \mathbf{r}') \cdot \mathbf{J}(\mathbf{r}'). \tag{2.9}$$

The real part of $\mathcal{P}$, i.e., $\Re\,[\mathcal{P}]$, represents the radiated power which is determined by the multipole moments $a_{l,m}^{(j)}$ via [14, 21, 30]

$$\Re\,[\mathcal{P}] = \frac{1}{2\eta_0}\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} l(l+1)|a_{l,m}^{(j)}|^2, \tag{2.10}$$

where $\eta_0 = \sqrt{\mu_0/\epsilon_0}$ is the free-space wave impedance.

On the other hand, the imaginary part of $\mathcal{P}$, i.e., $\Im\,[\mathcal{P}]$, corresponds to the energy-storage reactive power [7]. It can have a prescribed value, say, zero (as was shown in [32, 34]), which corresponds to a tuned antenna and is one of the solution constraints to be employed in the formulation to follow. We note that the reactive power can be expressed as

$$\Im\,[\mathcal{P}] = -\frac{1}{2}\int_V d\mathbf{r}\mathbf{J}^*(\mathbf{r})\cdot\int_V d\mathbf{r}'\bar{\mathbf{G}}_S(\mathbf{r},\mathbf{r}')\cdot\mathbf{J}(\mathbf{r}') \equiv -\frac{1}{2}(\mathbf{J},\widetilde{\mathbf{G}}_S\mathbf{J}), \tag{2.11}$$

where

$$\bar{\mathbf{G}}_S(\mathbf{r},\mathbf{r}') \equiv \Im\left[\bar{\mathbf{G}}(\mathbf{r},\mathbf{r}')\right] = \frac{1}{2i}\left[\bar{\mathbf{G}}(\mathbf{r},\mathbf{r}') - \bar{\mathbf{G}}^*(\mathbf{r},\mathbf{r}')\right], \tag{2.12}$$

and where we have introduced the linear mapping $\widetilde{\mathbf{G}}_S$ defined by (2.9) after the substitutions $\widetilde{\mathbf{G}} \to \widetilde{\mathbf{G}}_S$ and $\bar{\mathbf{G}} \to \bar{\mathbf{G}}_S$.

2.2. Source-to-multipole-moment mapping. To formulate the inverse problem for the cases described in (1.1) and (1.2) it is necessary to first have at our disposal the solution of the associated forward or radiation problem. To accomplish this, we note that, for these cases, the electric field $\mathbf{E}(\mathbf{r})$ generated by the most general source of support V can be represented, outside V, by the multipole expansion [14]

$$\mathbf{E}(\mathbf{r}) = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} a_{l,m}^{(j)}\mathbf{\Lambda}_{l,m}^{(j)}(\mathbf{r}), \quad \mathbf{r}\notin V, \tag{2.13}$$

where the complex-valued expansion coefficients $a_{l,m}^{(j)}$ are the multipole moments of the field, and where the multipole fields are

$$\mathbf{\Lambda}_{l,m}^{(j)}(\mathbf{r}) = \begin{cases} \mathbf{\nabla}\times[h_l^{(+)}(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], & j=1, \\ \\ ik_0 h_l^{(+)}(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & j=2, \end{cases} \tag{2.14}$$

where $\hat{\mathbf{r}} \equiv \mathbf{r}/r$; $h_l^{(+)}$ denotes the spherical Hankel function of the first kind and order l (as defined in [6]), corresponding to outgoing spherical waves in the far zone; $\mathbf{Y}_{l,m}$ is the vector spherical harmonic of degree l and order m (as defined in [14, equations (4.7) and (4.8)]); and $j=1$ and $j=2$ correspond to electric and magnetic multipole fields, respectively. The scalar spherical harmonics $Y_{l,m}(\hat{\mathbf{r}})$ and the vector spherical harmonics $\mathbf{Y}_{l,m}(\hat{\mathbf{r}})$ satisfy the analytic continuation properties $Y_{l,m}^*(\hat{\mathbf{r}}) = (-1)^m Y_{l,-m}(\hat{\mathbf{r}})$ and $\mathbf{Y}_{l,m}^*(\hat{\mathbf{r}}) = (-1)^{m+1}\mathbf{Y}_{l,-m}(\hat{\mathbf{r}})$ along with well-known orthogonality properties that can be found, e.g., in [14] (see [20] for further details). We will employ these properties in the following.

At this point it is important to note that the multipole moments $a^{(j)}_{l,m}$ are *uniquely* determined by the tangential component of the electric field $\mathbf{E}(\mathbf{r})$ on a sphere of radius $R > a$, in particular,

$$a^{(j)}_{l,m} = \begin{cases} -\frac{i}{l(l+1)k_0 h_l^{(+)}(k_0R)} \int \mathbf{Y}^*_{l,m}(\hat{\mathbf{r}}) \cdot \mathbf{E}(R\hat{\mathbf{r}})d\hat{\mathbf{r}}, & j = 1, \\ \frac{1}{l(l+1)k_0 V_l(k_0R)} \int \hat{\mathbf{r}} \times \mathbf{Y}^*_{l,m}(\hat{\mathbf{r}}) \cdot \mathbf{E}(R\hat{\mathbf{r}})d\hat{\mathbf{r}}, & j = 2. \end{cases} \tag{2.15}$$

(This follows by expanding (2.13) using [33]

$$\boldsymbol{\nabla} \times [\phi_l(r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})] = \hat{\mathbf{r}}\frac{il(l+1)}{r}\phi_l(r)Y_{l,m}(\hat{\mathbf{r}}) + \frac{1}{r}\frac{d}{dr}[r\phi_l(r)]\hat{\mathbf{r}} \times \mathbf{Y}_{l,m}(\hat{\mathbf{r}}), \tag{2.16}$$

and then invoking the orthogonality properties of the vector spherical harmonics and the associated vector functions $\hat{\mathbf{r}} \times \mathbf{Y}_{l,m}(\hat{\mathbf{r}})$.)

The electric and magnetic multipole moments, $a^{(1)}_{l,m}$ and $a^{(2)}_{l,m}$, respectively, are related to the current distribution $\mathbf{J}$ by

$$a^{(j)}_{l,m} = (\boldsymbol{\mathfrak{B}}^{(j)}_{l,m}, \mathbf{J}), \quad j = 1, 2; \tag{2.17}$$

i.e., they are the projections of the current distribution $\mathbf{J}$ onto the set of source-free vector fields $\boldsymbol{\mathfrak{B}}^{(j)}_{l,m}$ which need to be determined for the particular antenna background medium. For the special free-space case where $\mu(\mathbf{r})/\mu_0 = 1 = \epsilon(\mathbf{r})/\epsilon_0$ the latter fields are the familiar source-free multipole fields, in particular (cf. [19] and [11]),

$$\mathbf{B}^{(j)}_{l,m}(\mathbf{r}) \equiv \begin{cases} -\frac{\eta_0}{l(l+1)}\boldsymbol{\nabla} \times [j_l(k_0r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], & j = 1, \\ -i\frac{k_0\eta_0}{l(l+1)} j_l(k_0r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & j = 2, \end{cases} \tag{2.18}$$

where j_l is the spherical Bessel function of the first kind and order l (as defined in [6], for instance). On the other hand, it is shown in Appendix A that, for piecewise-constant radially symmetric backgrounds whose permittivity and permeability are given by (1.1) and (1.2),

$$\boldsymbol{\mathfrak{B}}^{(j)}_{l,m}(\mathbf{r}) \equiv \begin{cases} \frac{-\eta_0}{l(l+1)} F_l^{*(1)}(k_0a, ka, \epsilon_r, \mu_r)\boldsymbol{\nabla} \times [j_l(k^*r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], & j = 1, \\ \frac{-ik_0\eta_0}{l(l+1)} F_l^{*(2)}(k_0a, ka, \epsilon_r, \mu_r) j_l(k^*r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & j = 2, \end{cases} \tag{2.19}$$

where the substrate wavenumber $k = \omega\sqrt{\mu_s\epsilon_s}$, the relative permittivity $\epsilon_r = \epsilon_s/\epsilon_0$, the relative permeability $\mu_r = \mu_s/\mu_0$, and where we have defined the complex amplitudes

$$F_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r) \equiv \begin{cases} \frac{i/\left(k_0ka^2\right)}{(\epsilon_r/\mu_r)^{1/2} j_l(ka)V_l(k_0a) - h_l^{(+)}(k_0a)U_l(ka)}, & j = 1, \\ \frac{i\mu_r/\left(k_0ka^2\right)}{(\mu_r/\epsilon_r)^{1/2} j_l(ka)V_l(k_0a) - h_l^{(+)}(k_0a)U_l(ka)}, & j = 2, \end{cases} \tag{2.20}$$

where

$$U_l\,(\lambda a) \equiv U_l\,(\lambda r)|_{r=a} \equiv \left[\frac{dj_l(\lambda r)}{d\,(\lambda r)} + \frac{j_l(\lambda r)}{\lambda r}\right]\Bigg|_{r=a} \tag{2.21}$$

and

$$V_l\left(\lambda a\right) \equiv V_l\left(\lambda r\right)\big|_{r=a} \equiv \left[\frac{dh_l^{(+)}(\lambda r)}{d\left(\lambda r\right)} + \frac{h_l^{(+)}(\lambda r)}{\lambda r}\right]\Bigg|_{r=a} . \tag{2.22}$$

Because of the self-imposed restriction to the study of lossless substrates, the relative constitutive parameters μ_r and ϵ_r admit only real values. Consequently, the wavenumber k can assume only real values (positive for double-positive materials and negative for double-negative metamaterials) or purely imaginary values (for single-negative metamaterials). When k is purely imaginary, i.e., $k = i\alpha$, $\alpha \in \mathbb{R}$, the arguments of the spherical Bessel functions involving k in (2.20)–(2.22) are, accordingly, purely imaginary. In this case one notes that the regular spherical Bessel functions j_l and $h_l^{(+)}$ are replaced, respectively, with the modified spherical Bessel functions i_l and k_l such that (cf., for instance, [6])

$$j_l\left(ka\right) \equiv i^l i_l(\alpha a) \tag{2.23}$$

and

$$h_l^{(+)}\left(ka\right) \equiv -i^{-l} k_l(\alpha a). \tag{2.24}$$

(There shall be no confusion between the modified spherical Bessel functions i_l and k_l and the imaginary unit i and the wavenumber k since the latter do not carry a subscript.)

We draw the attention of the reader to the fact that $F_l^{(j)}$, $j = 1, 2$, represent the Mie amplitudes due to the scattering of a plane electromagnetic wave off a sphere of radius a and wavenumber k embedded in an infinite homogeneous medium of wavenumber k_0, $F_l^{(1)}$ being the amplitudes of the electric oscillations and $F_l^{(2)}$ those of the magnetic oscillations. This should not come as a surprise in view of the physics of the problem as well as the formulation itself.

3. Inverse source theory based on constrained optimization.

The inverse source problem of deducing the source $\mathbf{J}(\mathbf{r})$, confined within V from knowledge of the exterior field $\mathbf{E}(\mathbf{r})$, is seen from (2.13) to be equivalent to that of determining the source from knowledge of the multipole moments, i.e., to that of inverting (2.17). The respective inversion is addressed next via a generalization of the free-space optimization theory in [32] to nonhomogeneous backgrounds. Emphasis is given to the particular case of piecewise-constant radially symmetric backgrounds, but most of the derived expressions apply to more general cases including that of spherically symmetric backgrounds.

3.1. Minimum energy solution by constrained optimization.

We start by addressing the problem of determining the minimum energy source $\mathbf{J}_{ME}$ embedded in a substrate of volume V with *fixed* constitutive parameters ϵ_r, μ_r and generating a given exterior field. The problem can be cast as

$$\min_{\mathbf{J}\in S} \mathcal{E}\left(\mathbf{J}\right), \tag{3.1}$$

where

$$S \equiv \left\{\mathbf{J} \in L^2\left(V; \mathbb{C}^3\right) : a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}, \mathbf{J}) = 0\right\}. \tag{3.2}$$

Note that the constraint set S is convex; also, the objective functional $\mathcal{E}$ is coercive and strictly convex.

If a minimizer $\mathbf{J}_{ME}$ exists, then its uniqueness and global minimality are ensured by the strict convexity of $\mathcal{E}$ and the convexity of S [4]. But what guarantees the existence of at least one such minimizer? We note that, since S is closed and since $\mathcal{E}$ is weakly sequentially lower semicontinuous, problem (3.1), (3.2) admits only one global solution.

It is well known [4] that difficulties related to the definition of linearity of the Fréchet differentiation operator would be encountered when f maps a complex Banach space (in our case $L^2(V;\mathbb{C}^3)$) into a real Banach space (in our case $\mathbb{R}$). This is in particular the case for $\mathcal{E}$ and $(\mathbf{J}, \widetilde{\mathbf{G}}_S\mathbf{J})$. In such cases one considers $L^2(V;\mathbb{C}^3)$ as a Hilbert space over $\mathbb{R}$ instead of $\mathbb{C}$ [4]. It is then easy to show that the Fréchet derivatives of $\mathcal{E}$ and the constraints are given by $\boldsymbol{\nabla}_\mathbf{J}\mathcal{E}(\mathbf{J}) = 2\mathbf{J}$ and that $\boldsymbol{\nabla}_\mathbf{J}\,[a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}, \mathbf{J})] = -\mathfrak{B}_{l,m}^{(j)}$, respectively. Due to the continuity of these derivatives and the fact that $\boldsymbol{\nabla}_\mathbf{J}\,[a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}, \mathbf{J})]|_{\mathbf{J}=\mathbf{J}_{ME}}$ maps $L^2(V;\mathbb{C}^3)$ onto $\mathbb{C}$, there exist [23] Lagrange multipliers $c_{l,m}^{(j)} \in \mathbb{C}$ such that the generalized Lagrangian

$$\mathcal{L}\left(\mathbf{J}, c_{l,m}^{(j)}\right) \equiv \mathcal{E} + 2\Re\left[\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)}\left(a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}\,,\, \mathbf{J})\right)\right] \tag{3.3}$$

is stationary at $\mathbf{J}_{ME}$.

To compute the solution we require that

$$\delta\mathcal{L} = 2\Re\left[\left(\delta\mathbf{J}\,,\,\mathbf{J} - \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)*}\mathfrak{B}_{l,m}^{(j)}\right)\right] = 0. \tag{3.4}$$

From the Du Bois-Raymond lemma, (3.4), and the forward mapping relations (2.17), (2.19), (2.20), one finds that for piecewise-constant radially symmetric backgrounds, the minimum-energy source is given by

$$\mathbf{J}_{ME}(\mathbf{r}) = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} \frac{a_{l,m}^{(j)}}{\left[\sigma_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r)\right]^2}\mathfrak{B}_{l,m}^{(j)}(\mathbf{r}), \tag{3.5}$$

where we have introduced the positive-definite "singular values"

$$[\sigma_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r)]^2 \equiv (\mathfrak{B}_{l,m}^{(j)}, \mathfrak{B}_{l,m}^{(j)}), \tag{3.6}$$

specifically,

$$\left[\sigma_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r)\right]^2 = |F_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r)|^2\left[\kappa_l^{(j)}(k_0a, ka)\right]^2, \tag{3.7}$$

where

$$\left[\kappa_l^{(j)}(k_0a, ka)\right]^2 \equiv \begin{cases} \eta_0^2\int_0^a dr\left[|j_l(kr)|^2 + \frac{|kr|^2}{l(l+1)}\,|U_l(kr)|^2\right], & j = 1, \\ \\ \frac{\eta_0^2 k_0{}^2}{l(l+1)}\int_0^a dr\, r^2\,|j_l(kr)|^2\,, & j = 2. \end{cases} \tag{3.8}$$

For real k^2 the integral associated with the $j = 2$ case is calculable through the use of the second Lommel integral (see, for instance, [6]) and the recurrence relations of the Bessel functions along lines similar to those employed in [31] to evaluate similar inner products. Afterwards, the recurrence relations are also used to express the integral associated with the $j = 1$ case in terms of the calculated integral associated with the $j = 2$ case. Consequently, equations (3.8) reduce to

$$\left[\kappa_l^{(j)}\right]^2 = \begin{cases} \frac{\eta_0^2 a|ka|^2}{l(l+1)(2l+1)}\left[(l+1)\gamma_{l-1}^2(ka) + \ l\gamma_{l+1}^2(ka)\right], & j=1, \\ \frac{\eta_0^2 a(k_0a)^2}{l(l+1)}\gamma_l^2(ka), \quad j=2, \end{cases} \tag{3.9}$$

where we have introduced the unitless quantity (cf. [31, equation (17)])

$$\begin{aligned} \gamma_l^2(ka) &\equiv \frac{1}{a^3}\int_0^a dr\, r^2 j_l^2(kr) \\ &= \frac{1}{2}\left[j_l^2(ka) - j_{l-1}(ka)j_{l+1}(ka)\right]. \end{aligned} \tag{3.10}$$

For $k = i\alpha$, $\alpha \in \mathbb{R}$, as is the case for single-negative metamaterials, one uses definition (2.23) to express (3.9) and (3.10) in terms of i_l. (Note that the lone appearance of the size parameter a in (3.9), i.e., its appearance decoupled from the wavenumbers, is a direct consequence of the fact that the multipole moments $a_{l,m}^{(j)}$ are dimensioned quantities. It is, as well, a reminder of the boundedness of the enclosing volume V, i.e., of the embedding sphere of substrate material.)

Furthermore, the minimum source energy

$$\mathcal{E}_{ME} \equiv (\mathbf{J}_{ME}, \mathbf{J}_{ME}) = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} \frac{|a_{l,m}^{(j)}|^2}{[\sigma_l^{(j)}]^2}. \tag{3.11}$$

As expected, these developments reduce, for $\epsilon_r = 1 = \mu_r$, to the free-space result ((13) and (14) in [32]) since $F_l^{(j)}(k_0a, ka, \epsilon_r, \mu_r) = 1$; that is, the free-space minimum energy solution is given by (3.5) with $\mathfrak{B}_{l,m}^{(j)}$ given by (2.18) and $[\sigma_l^{(j)}]^2$ substituted by $[\kappa_l^{(j)}(k_0a = ka)]^2$.

3.2. Minimum energy source having zero reactive power. Next we consider the constrained optimization problem of minimizing the functional energy of the source subject to the additional constraint that the reactive power of the source has a prescribed value. The results for this problem will be elaborated next for the particular and important case of zero reactive power, i.e., $\Im[\mathcal{P}] = 0$. This corresponds to the minimizing of the antenna currents (the physical resources) while simultaneously enforcing perfect antenna reactance tuning inside the antenna.

The problem can be cast as

$$\min_{\mathbf{J}\in X} \mathcal{E}(\mathbf{J}), \tag{3.12}$$

where

$$X \equiv \left\{\mathbf{J} \in L^2\left(V; \mathbb{C}^3\right) : a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}, \mathbf{J}) \ = 0, \ (\mathbf{J}, \widetilde{\mathbf{G}}_S\mathbf{J}) = 0\right\}. \tag{3.13}$$

The constraint set X is closed, unbounded, and nonconvex. Its nonconvexity stems from that of the newly introduced constraint $(\mathbf{J}, \widetilde{\mathbf{G}}_S\mathbf{J}) = 0$. The set X is assumed to be nonempty. (If it turns out to be empty, this would mean that it is not possible for an antenna having a substrate medium of constitutive parameters ϵ_r, μ_r to produce the prescribed external field and at the same time have a vanishing reactive power.)

It is clear that problem (3.12), (3.13) is an inherently difficult nonconvex programming problem. Not only do we seek to minimize an objective functional under nonconvex functional constraints, but we also have to do that on an unbounded set. Proving, for instance, the existence of a solution to problem (3.12), (3.13) would have been easier if X were convex, but it is straightforward to show that the only way for X to become convex is to have $\Re[(\mathbf{J}_1, \widetilde{\mathbf{G}}_S\mathbf{J}_2)] \leq 0$ for all $\mathbf{J}_1, \mathbf{J}_2 \in L^2\left(V; \mathbb{C}^3\right)$. This would amount to imposing a new constraint which appears not to correspond to anything meaningful, physically speaking.

Now let us try to establish the existence of a solution to problem (3.12), (3.13) in the absence of the convexity and boundedness of the constraint set X. Since X is a closed subset of a normed vector space and since $\mathcal{E}$ is a coercive functional, there exist [23] $\mathbf{J}_0 \in X$ and $\Gamma > 0$ such that

$$\inf_{\mathbf{J}\in X} \mathcal{E}\left(\mathbf{J}\right) = \inf\left\{\mathcal{E}\left(\mathbf{J}\right) : \mathbf{J} \in X \cap \overline{B_\Gamma\left(\mathbf{J}_0\right)}\right\}, \tag{3.14}$$

where $\overline{B_\Gamma\left(\mathbf{J}_0\right)}$ is the closed (and bounded) ball of radius Γ and center $\mathbf{J}_0$. This is a powerful result. What this tells us is that minimizing $\mathcal{E}$ over the unbounded set X can be reduced to minimizing $\mathcal{E}$ over a bounded subset in X that could be much smaller than X. All that remains to complete the proof of existence of a solution to problem (3.12), (3.13) is to demonstrate the existence of a solution to the auxiliary problem

$$\min_{\mathbf{J}\in X\cap\overline{B_\Gamma(\mathbf{J}_0)}} \mathcal{E}\left(\mathbf{J}\right). \tag{3.15}$$

A useful variant of the generalized Weierstrass theorem stipulates that for a weakly sequentially lower semicontinuous functional defined on a weakly sequentially compact subset of a Hilbert space there exists at least one solution to the minimization problem [4]. But we have already shown that $\mathcal{E}$ is a weakly sequentially lower semicontinuous functional (see the discussion of problem (3.1), (3.2)). Consequently, the existence of a solution to problem (3.15), (3.13) depends entirely on the demonstration that $X \cap \overline{B_\Gamma\left(\mathbf{J}_0\right)}$ is a weakly sequentially compact subset. But this, too, is true because any bounded subset of a reflexive Banach space (e.g., a Hilbert space) is also weakly sequentially compact [8]. Hence, assuming that $X \cap \overline{B_\Gamma\left(\mathbf{J}_0\right)}$ is nonempty, we are, from the preceding discussion, in a position to affirm the existence of at least one global minimizer $\mathbf{J}_{\mathcal{E},\mathcal{P}} \in X \cap \overline{B_\Gamma\left(\mathbf{J}_0\right)}$ for the auxiliary problem (3.15), (3.13). However, by virtue of (3.14), this point $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ is also the sought solution of problem (3.12), (3.13), which completes our proof.

Unfortunately, though, we have yet to guarantee the uniqueness of this solution or even write down a minimality condition that would yield this solution. We shall now focus on trying to write down a necessary minimality condition whose solution would, at least in principle, yield a candidate $\mathbf{J}_{\mathcal{E},\mathcal{P}}$. The Fréchet derivative of the nonconvex constraints is given by $\boldsymbol{\nabla}_{\mathbf{J}}(\mathbf{J}, \widetilde{\mathbf{G}}_S\mathbf{J}) = 2\widetilde{\mathbf{G}}_S\mathbf{J}$. Now, let $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ be a minimizer. In view of the noted Fréchet differentiability of the objective functional and the constraints, the continuity of their Fréchet derivatives, and the fact that $\boldsymbol{\nabla}_{\mathbf{J}}\,[a_{l,m}^{(j)} - (\mathfrak{B}_{l,m}^{(j)}, \mathbf{J})]|_{\mathbf{J}=\mathbf{J}_{\mathcal{E},\mathcal{P}}}$ is surjective and the range of $\boldsymbol{\nabla}_{\mathbf{J}}(\mathbf{J}, \widetilde{\mathbf{G}}_S\mathbf{J})|_{\mathbf{J}=\mathbf{J}_{\mathcal{E},\mathcal{P}}}$ is closed, there exist Lagrange multipliers $\chi \in \mathbb{R}$ and $c_{l,m}^{(j)} \in \mathbb{C}$ such that [23]

$$\Re\left[\left(\boldsymbol{\nabla}_{\mathbf{J}}\mathcal{L}\left(\mathbf{J}_{\mathcal{E},\mathcal{P}},\chi,c_{l,m}^{(j)}\right)\ ,\ \mathbf{J}-\mathbf{J}_{\mathcal{E},\mathcal{P}}\right)\right]\geq 0\quad\forall\,\mathbf{J}\in L^{2}\left(V;\mathbb{C}^{3}\right), \tag{3.16}$$

where the generalized Lagrangian functional is given by

$$\begin{aligned}\mathcal{L}\left(\mathbf{J},\chi,c_{l,m}^{(j)}\right)&\equiv\mathcal{E}\left(\mathbf{J}\right)+\chi\left(\mathbf{J},\widetilde{\mathbf{G}}_{S}\mathbf{J}\right)\\&\quad+2\Re\left\{\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}c_{l,m}^{(j)}\left[a_{l,m}^{(j)}-(\mathfrak{B}_{l,m}^{(j)},\mathbf{J})\right]\right\}.\end{aligned} \tag{3.17}$$

Condition (3.16), (3.17) reduces to

$$\Re\left[\left(\mathbf{J}_{\mathcal{E},\mathcal{P}}+\chi\widetilde{\mathbf{G}}_{S}\mathbf{J}_{\mathcal{E},\mathcal{P}}-\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}c_{l,m}^{(j)}\mathfrak{B}_{l,m}^{(j)}\ ,\ \mathbf{J}-\mathbf{J}_{\mathcal{E},\mathcal{P}}\right)\right]\geq 0\quad\forall\,\mathbf{J}\in L^{2}\left(V;\mathbb{C}^{3}\right). \tag{3.18}$$

According to (3.18), to determine $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ one needs to solve an infinite number of equations with an infinite number of unknowns. That, of course, is not the case in practical situations. For any real problem the radiation emitted by the source has a maximum multipolarity $l_{\max}\sim ka$ $(<\infty)$. Thus for real problems one would need to solve $2l_{\max}\left(l_{\max}+2\right)+4$ integral equations with $2l_{\max}\left(l_{\max}+2\right)+4$ unknowns. By all standards this is a tedious task, even for small values of $l_{\max}$. One should try to find a more clever way of determining what the solution is. For instance, one could resort to numerical techniques and algorithms available in the literature (see, e.g., [44] and the references therein). In what follows we plan on adopting a similar approach that combines analytical and numerical methods.

We shall *assume* that $X\neq\emptyset$ and adopt partly analytical, partly numerical strategies to find a candidate $\mathbf{J}_{\mathcal{E},\mathcal{P}}$, which we proved exists, without having to solve a large number of complicated equations. The "hybrid" approach below is very much in line with the spirit of those adopted for this kind of problem. We shall also try to explore some of the properties of the solution. Once a feasible point $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ is found by means of the technique below, one would substitute it into the derived minimality conditions to check whether it satisfies these conditions.

Now let $\mathcal{L}$ be the generalized Lagrangian defined as

$$\begin{aligned}\mathcal{L}\left(\mathbf{J},\chi,c_{l,m}^{(j)}\right)&\equiv\mathcal{E}\left(\mathbf{J}\right)+\chi\left\{(\mathbf{J},\widetilde{\mathbf{G}}_{S}\mathbf{J})+2\Im[\mathcal{P}]\right\}\\&\quad+2\Re\left\{\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}c_{l,m}^{(j)}\left[a_{l,m}^{(j)}-(\mathfrak{B}_{l,m}^{(j)},\mathbf{J})\right]\right\},\end{aligned} \tag{3.19}$$

wherein the constraint on the reactive power is now written in such a way that it permits the latter to have an arbitrary value $\Im[P]$ that is not necessarily zero.

The first variation of the last term in (3.19) is found from (2.11) and (2.12) to be

$$\chi\delta(\mathbf{J},\widetilde{\mathbf{G}}_{S}\mathbf{J})=2\Re\left[\chi(\delta\mathbf{J},\widetilde{\mathbf{G}}_{S}\mathbf{J})\right]. \tag{3.20}$$

It follows from (3.3), (3.4), and (3.20) that the first variation of the Lagrangian in (3.19) is

$$\delta\mathcal{L} = 2\Re\left[(\delta\mathbf{J},\mathbf{J}) + \chi(\delta\mathbf{J},\widetilde{\mathbf{G}}_S\mathbf{J}) - \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)*}(\delta\mathbf{J},\mathfrak{B}_{l,m}^{(j)})\right]. \tag{3.21}$$

By equating the variation in (3.21) to zero, one deduces that the sought solution, to be denoted as $\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r})$, must obey, within its support V, the relation

$$\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) + \chi\widetilde{\mathbf{G}}_S\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)*}\mathfrak{B}_{l,m}^{(j)}(\mathbf{r}). \tag{3.22}$$

If $\chi = 0$, then this approach coincides with the one given earlier, leading to the minimum energy source in (3.5) (in such a situation, that source generates zero reactive power), while for the more general case $\chi \neq 0$ the two formulations (and their solutions) differ. However, we note that for certain peculiar constitutive-parameter values the constraint is not active and therefore $\chi = 0$. In that peculiar case the minimum-energy sources are intrinsically resonant.

By letting the vector wave equation operator $(\nabla\times\nabla\times - (k^*)^2) = (\nabla\times\nabla\times - k^2)$ (the equality stems from the requirement that the substrate be lossless) act on both sides of (3.22) and with the aid of the fact that the fields $\mathfrak{B}_{l,m}^{(j)}$ are solutions of the homogeneous wave equation associated to the same operator, one concludes that the source $\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r})$ obeys the homogeneous wave equation

$$\nabla\times\nabla\times\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) - K^2\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) = \mathbf{0} \tag{3.23}$$

in the interior of the source region V; the quantity K which appears in (3.23) is a modified wavenumber defined by

$$K^2 \equiv k^2 - \chi\mu_s\omega. \tag{3.24}$$

(Note that K quickly becomes purely imaginary as χ becomes large and positive.)

Now, the most general source that is confined within the spherical source volume V and is a solution of (3.23) in the interior of V must admit the representation

$$\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} v_{l,m}^{(j)}\mathfrak{D}_{l,m}^{(j)}(\mathbf{r}), \tag{3.25}$$

where $v_{l,m}^{(j)}$ are expansion coefficients that need to be determined (for the constraints of the problem) and where

$$\mathfrak{D}_{l,m}^{(j)}(\mathbf{r}) = \begin{cases} -\frac{\eta_0}{l(l+1)}\nabla\times[j_l(Kr)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], & j=1, \\ -\frac{i\eta_0 K}{l(l+1)} j_l(Kr)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & j=2. \end{cases} \tag{3.26}$$

From the formal similarity of $\mathfrak{B}_{l,m}^{(j)}$ and $\mathfrak{D}_{l,m}^{(j)}$ (cf. (2.19)) it follows at once from (3.6), (3.7), (3.9), and (3.10) that the inner product

$$\left(\mathfrak{D}_{l,m}^{(j)},\mathfrak{D}_{l,m}^{(j)}\right) = p^{(j)}\left[\kappa_l^{(j)}(k_0a,Ka)\right]^2, \tag{3.27}$$

where

$$p^{(j)} = \begin{cases} 1, \quad j = 1, \\ \\ |K|^2/k_0^2, \quad j = 2. \end{cases} \tag{3.28}$$

By substituting from (3.25) and (3.26) into (2.17) while using well-known orthogonality properties of the vector spherical harmonics $\mathbf{Y}_{l,m}(\hat{\mathbf{r}})$ and the associated vector functions $\hat{\mathbf{r}} \times \mathbf{Y}_{l,m}(\hat{\mathbf{r}})$ one obtains

$$\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) = \sum_{j=1}^{2} \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \frac{a_{l,m}^{(j)}}{(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)})} \mathfrak{D}_{l,m}^{(j)}(\mathbf{r}), \tag{3.29}$$

where

$$(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}) = \begin{cases} \eta_0^2 F_l^{(1)} \int_0^a dr \left[j_l(kr) j_l(Kr) + \frac{kKr^2}{l(l+1)} U_l\,(kr)\, U_l\,(Kr) \right], \\ \hfill j = 1, \\ \\ \eta_0^2 F_l^{(2)} \frac{k_0 K}{l(l+1)} \int_0^a drr^2 j_l(kr) j_l(Kr), \quad j = 2. \end{cases} \tag{3.30}$$

Similarly to the integrals in (3.8), the integral associated with the $j = 2$ case in (3.30) is calculable through the use of the first Lommel integral (cf., for instance, [6]). The above inner product takes the form

$$(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}) = \begin{cases} \frac{\eta_0^2 kKa^3 F_l^{(1)}}{l(l+1)(2l+1)} \left[(l+1)\psi_{l-1}\,(ka, Ka) + \; l\psi_{l+1}\,(ka, Ka) \right], \\ \hfill j = 1, \\ \\ \frac{\eta_0^2 k_0 Ka^3 F_l^{(2)}}{l(l+1)} \psi_l\,(ka, Ka)\,, \quad j = 2, \end{cases} \tag{3.31}$$

where we have introduced the unitless quantity

$$\begin{aligned} \psi_l\,(ka, Ka) &\equiv \frac{1}{a^3} \int_0^a drr^2 j_l(kr) j_l(Kr) \\ &= \frac{1}{a\,(k^2 - K^2)} \left[Kj_l(ka) j_{l-1}(Ka) - kj_{l-1}(ka) j_l(Ka) \right]. \end{aligned} \tag{3.32}$$

(Note that (3.32) is valid only for $k \neq K$, i.e., for $\chi \neq 0$. The case $k = K$, i.e., for $\chi = 0$, has already been discussed.)

The source energy corresponding to (3.29) is of the form

$$\mathcal{E}_{\mathcal{E},\mathcal{P}} \equiv (\mathbf{J}_{\mathcal{E},\mathcal{P}}, \mathbf{J}_{\mathcal{E},\mathcal{P}}) = \sum_{j=1}^{2} \sum_{l=1}^{\infty} \sum_{m=-l}^{l} \frac{(\mathfrak{D}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)})}{|(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)})|^2} |a_{l,m}^{(j)}|^2. \tag{3.33}$$

Note that the above results (3.29) and (3.33) do not assume any particular value for $\Im\,[\mathcal{P}]$.

We need to incorporate the reactive power constraint, i.e., (2.11), which defines the value of the remaining Lagrange multiplier χ. Since the desired reactive power is specified to be zero, the problem now is to find an expression for the reactive power in terms of χ from which one can deduce the value of χ which minimizes the source energy under the constraint $\Im\left[\mathcal{P}\right] = 0$. This value of χ will be called χ_0. A number of partly analytical, partly numerical strategies can be implemented to accomplish this step.

One such approach, which generalizes the development for the free-space case in [32], consists of determining the field $\mathbf{E}(\mathbf{r})$ generated by the source $\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r})$ in the interior of the source region V. In particular, after evaluating the field, one can compute the interaction power via (2.8) and (2.9) and require that its imaginary part vanish. In particular, plotting $\Im\left[\mathcal{P}\right]$ and $\mathcal{E}_{\mathcal{E},\mathcal{P}}$ versus χ one can finally select the value of χ which yields minimum $\mathcal{E}_{\mathcal{E},\mathcal{P}}$ out of all values of χ for which $\Im\left[\mathcal{P}\right] = 0$. We adopt this approach next.

By rewriting (3.23) as

$$\left(\boldsymbol{\nabla} \times \boldsymbol{\nabla} \times -k^2\right)\left[\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) - i\chi\mathbf{E}(\mathbf{r})\right] = \mathbf{0}, \tag{3.34}$$

where we have borrowed from (2.3), one concludes that the field $\mathbf{E}(\mathbf{r})$ must admit in the interior of the source region V an expansion of the form

$$\mathbf{E}(\mathbf{r}) = \frac{1}{i\chi}\left[\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r}) + \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} u_{l,m}^{(j)}\mathfrak{B}_{l,m}^{(j)}(\mathbf{r})\right], \quad \mathbf{r} \in V, \tag{3.35}$$

where the expansion coefficients $u_{l,m}^{(j)}$ need to be determined taking into account continuity of the tangential components of the field on the boundary $\partial V \equiv \{\mathbf{r} \in \mathbb{R}^3 : r = a\}$ of V. Continuing with this idea, it is not hard to show from these developments, and by straightforward generalization of the discussion of the free-space version of the problem in [32], equations (30)–(42), that the complex interaction power can be expressed as

$$\mathcal{P} = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} q_l^{(j)}|a_{l,m}^{(j)}|^2, \tag{3.36}$$

where

$$q_l^{(j)} = \frac{i}{2\chi}\left[\frac{\left(\mathfrak{D}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}\right)}{\left|\left(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}\right)\right|^2} + \frac{u_{l,m}^{(j)}}{a_{l,m}^{(j)}}\right], \tag{3.37}$$

where the quantity $u_{l,m}^{(j)}/a_{l,m}^{(j)}$ is given by

$$\frac{u_{l,m}^{(j)}}{a_{l,m}^{(j)}} = \begin{cases} \frac{1}{F_l^{*(1)}k^*U_l(k^*a)}\left[-i\frac{k_0}{\eta_0}\chi l(l+1)V_l(k_0a) - \frac{KU_l(Ka)}{\left(\mathfrak{B}_{l,m}^{(1)}, \mathfrak{D}_{l,m}^{(1)}\right)}\right], \\ \hfill j = 1, \\ \\ \frac{1}{F_l^{*(2)}k_0j_l(k^*a)}\left[-i\frac{k_0}{\eta_0}\chi l(l+1)h_l^{(+)}(k_0a) - \frac{Kj_l(Ka)}{\left(\mathfrak{B}_{l,m}^{(2)}, \mathfrak{D}_{l,m}^{(2)}\right)}\right], \\ \hfill j = 2, \end{cases} \tag{3.38}$$

where the radial functions U_l and V_l have already been defined in (2.21) and (2.22), respectively.

Thus the reactive power of the source $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ is given by

$$\Im\left[\mathcal{P}\right] = \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} g_l^{(j)} |a_{l,m}^{(j)}|^2, \tag{3.39}$$

where

$$g_l^{(j)} = \frac{1}{2\chi}\left\{ \frac{\left(\mathfrak{D}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}\right)}{\left|\left(\mathfrak{B}_{l,m}^{(j)}, \mathfrak{D}_{l,m}^{(j)}\right)\right|^2} + \Re\left[\frac{u_{l,m}^{(j)}}{a_{l,m}^{(j)}}\right]\right\}. \tag{3.40}$$

By taking the real part of the complex interaction power, as given by (3.36)–(3.38), one also recovers (2.10), which is the well-known expression for the radiated power in terms of the multipole moments.

Equations (3.38)–(3.40) relate χ directly to $\Im\left[\mathcal{P}\right]$, as desired. For a certain problem, where $a_{l,m}^{(j)}$ and $\Im\left[\mathcal{P}\right]$ are given, one can compute the values of χ for which $\Im\left[\mathcal{P}\right] = 0$ by using these expressions, and pick, out of those values, the one which minimizes the functional energy in (3.33). By substituting that value of χ (i.e., χ_0) into (3.24), (3.26), and (3.29) one arrives at the desired solution.

Let

$$\Xi \equiv \left\{\chi \in \mathbb{R} : \Im\left[\mathcal{P}\left(\chi\right)\right] = 0\right\}. \tag{3.41}$$

It is found, numerically, that (see section 4) the minimum source energy is achieved for the value of χ that is closest to $\chi = 0$, i.e.,

$$|\chi_0| = \inf_{\chi\in\Xi}\left\{|\chi|\right\}. \tag{3.42}$$

It appears only natural to assume that an increase in the source energy from $\mathcal{E}_{ME}$ should correspond to χ_0 (and any other value of $\chi \in \Xi$ for that matter). This would be understood, intuitively, as a cost that one would have to pay to realize a tuned antenna. The numerical simulations suggest that this, in fact, is the case: Substituting any nonzero value $\chi \in \Xi$ in the expression for $\mathcal{E}_{\mathcal{E},\mathcal{P}}$ yields a value that is larger than $\mathcal{E}_{\mathcal{E},\mathcal{P}}|_{\chi=0} = \mathcal{E}_{ME}$. Does this mean that $\mathcal{E}_{ME}$ is a lower bound of $\mathcal{E}_{\mathcal{E},\mathcal{P}}$ and that $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ is a global minimum? Before we examine this question we note that the above observations remind us of (3.14). Indeed, a convenient way of viewing these observations is to think of the origin of the sphere $\overline{B_\Gamma\left(\mathbf{J}_0\right)} \ni \mathbf{J}_{\mathcal{E},\mathcal{P}}$ as the point $\mathbf{J}_0 = \mathbf{J}_{ME}$ and to think of its radius as $\Gamma \geq |\chi_0|$.

Supposing that $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ corresponds to a feasible point, let us try to derive a condition for it to be a global minimum. By definition, $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ is said to be a global minimum when

$$\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi\right) \geq \mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right) \quad \forall\chi \in \Xi. \tag{3.43}$$

If the inequality is strict, then the global minimum is also unique.

It follows from (2.17), (2.11), and (3.22) that

$$\begin{aligned}\mathcal{E}_{\mathcal{E},\mathcal{P}} - 2\chi\Im[\mathcal{P}] &= \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)*}(\mathbf{J}_{\mathcal{E},\mathcal{P}}, \mathfrak{B}_{l,m}^{(j)})\\ &= \sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l} c_{l,m}^{(j)*}a_{l,m}^{(j)*}.\end{aligned} \tag{3.44}$$

Thus if we require $\Im\left[\mathcal{P}\left(\chi\right)\right]=0$, (3.44) yields

$$\mathcal{E}_{\mathcal{E},\mathcal{P}}=\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}c_{l,m}^{(j)*}a_{l,m}^{(j)*}. \tag{3.45}$$

Furthermore, by projecting both sides of (3.22) onto the functions $\mathfrak{B}_{l,m}^{(j)}$ while recalling (2.17) and (3.6), one obtains

$$a_{l,m}^{(j)}+\chi(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathbf{J}_{\mathcal{E},\mathcal{P}})=c_{l,m}^{(j)*}[\sigma_l^{(j)}]^2. \tag{3.46}$$

By substituting from this result into (3.45), one obtains

$$\mathcal{E}_{\mathcal{E},\mathcal{P}}=\mathcal{E}_{ME}+\chi\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\frac{(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathbf{J}_{\mathcal{E},\mathcal{P}})a_{l,m}^{(j)*}}{[\sigma_l^{(j)}]^2}, \tag{3.47}$$

where $\chi\in\Xi$. Upon substituting $\mathbf{J}_{\mathcal{E},\mathcal{P}}(\mathbf{r})$ from (3.29) into (3.47) and using standard orthogonality properties of the spherical harmonics, one obtains

$$\mathcal{E}_{\mathcal{E},\mathcal{P}}=\mathcal{E}_{ME}+\chi\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\frac{(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathfrak{D}_{l,m}^{(j)}\left(\chi\right))}{(\mathfrak{B}_{l,m}^{(j)},\mathfrak{D}_{l,m}^{(j)}\left(\chi\right))}\frac{|a_{l,m}^{(j)}|^2}{[\sigma_l^{(j)}]^2}. \tag{3.48}$$

Expression (3.48) for the source energy directly assumes that $\Im\left[\mathcal{P}\right]=0$, while expression (3.33) holds for any value of the reactive power $\Im\left[\mathcal{P}\right]$.

It follows from (3.43) and (3.48) that the condition for $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ to be a global minimum is given by

$$\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\left\{\chi\frac{(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathfrak{D}_{l,m}^{(j)}\left(\chi\right))}{(\mathfrak{B}_{l,m}^{(j)},\mathfrak{D}_{l,m}^{(j)}\left(\chi\right))}-\chi_0\frac{(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathfrak{D}_{l,m}^{(j)}\left(\chi_0\right))}{(\mathfrak{B}_{l,m}^{(j)},\mathfrak{D}_{l,m}^{(j)}\left(\chi_0\right))}\right\}\frac{|a_{l,m}^{(j)}|^2}{[\sigma_l^{(j)}]^2}\geq 0 \tag{3.49}$$

for any value of $\chi\in\Xi$. Condition (3.49) is a necessary and sufficient condition for $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ to be a global minimum. The way it should be used is as follows. For a given substrate, solve $\Im\left[\mathcal{P}\left(\chi\right)\right]=0$ for χ (where $\Im\left[\mathcal{P}\left(\chi\right)\right]$ is given by (3.38)–(3.40)). If condition (3.49) is satisfied for *all* values of $\chi\in\Xi$, then $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ is a global minimum. If it is not satisfied for at least one value of $\chi\in\Xi$, then $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\right)$ is not a global minimum (but it may still be a local minimum).

Condition (3.49) was written based on the presumption that $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0\neq 0\right)$ was a global minimum. For $\mathcal{E}_{\mathcal{E},\mathcal{P}}\left(\chi_0=0\right)=\mathcal{E}_{ME}$, condition (3.49) reduces to

$$\chi\sum_{j=1}^{2}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\frac{(\mathfrak{B}_{l,m}^{(j)},\widetilde{\mathbf{G}}_S\mathfrak{D}_{l,m}^{(j)})}{(\mathfrak{B}_{l,m}^{(j)},\mathfrak{D}_{l,m}^{(j)})}\frac{|a_{l,m}^{(j)}|^2}{[\sigma_l^{(j)}]^2}\geq 0\quad\forall\chi\in\Xi. \tag{3.50}$$

4. Computer simulation study. The previous theory and algorithms are applied next to elucidate the effect of the antenna-embedding medium on radiation performance for two classes of antennas: electrically small and larger (resonant) antennas.

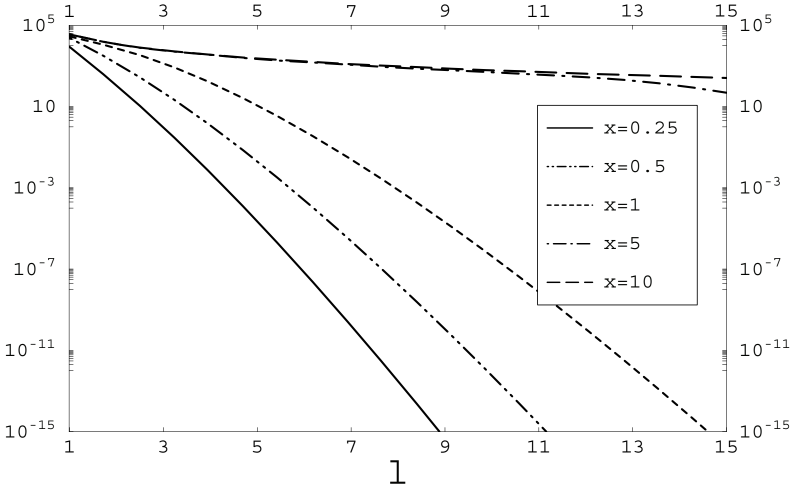


FIG. 4.1. *Free-space singular values* $[\sigma_l^{(1)}\ (x = x_0,\ \epsilon_r = +1 = \mu_r)]^2$ *versus* l *for a few representative values of* $x_0 \equiv k_0 a/\pi$. *(The unit of the singular values is* $V^2 m/A^2$*.)*

4.1. Minimum energy sources. It follows from (3.11) that, generally, the larger the singular values $[\sigma_l^{(j)}]^2$, the smaller the minimum source energy $\mathcal{E}_{ME}$ required for the launching of a given radiation pattern with a source of a given size. The singular values $[\sigma_l^{(j)}(k_0 = k,\ \epsilon_r = +1 = \mu_r)]^2$ correspond to the source in free space, that is, without the substrate. Thus, the larger the singular values $[\sigma_l^{(j)}]^2$ for a given substrate wavenumber ka relative to the corresponding free-space values, the greater the associated enhancement, due to the substrate, of radiation of the lth multipole order field with given resources. It is thus important to understand the dependence of the singular values $[\sigma_l^{(j)}(k_0 a, ka, \epsilon_r, \mu_r)]^2$ on $k_0 a$, ka, ϵ_r, μ_r, and l, for both the electric ($j = 1$) and the magnetic ($j = 2$) cases. Large singular values, such as resonances or peaks in the plots of the singular values versus these variables, will indicate enhanced radiation for such operational modes or conditions, with the given resources. This aspect is investigated numerically next.

Before engaging in the numerical illustrations we make some remarks: (1) the multipolarity l is handled in the plots as a continuous variable to facilitate understanding of the curves, yet the meaningful results correspond solely to the discrete values of l; (2) in the simulations the size parameter (radius) a of the antenna including the substrate has been set to unity, i.e., $a = 1$ meter; and (3) in the plots and associated discussion we consider the normalized wavenumbers defined by $x \equiv ka/\pi$ and $x_0 \equiv k_0 a/\pi$. The normalized wavenumber x represents the wavenumber of the field in the material, hence, the effective electric size in the material, while the normalized wavenumber x_0 measures the respective size in free space.

4.1.1. Behavior of the singular values $[\sigma_l^{(j)}]^2$. Figure 4.1 shows, for different antenna sizes, the free-space singular values $[\sigma_l^{(1)}(x_0 = x, , \epsilon_r = 1 = \mu_r)]^2$. No local maxima or resonances are seen for the free-space cases; in particular, in those cases the singular value spectrum decays exponentially. Figure 4.2 shows, for an antenna whose size corresponds to that of a quarter-wave antenna in free space ($x_0 = 1/4$), plots of the normalized electric singular values $[\varrho_l^{(1)}(x_0, x, \epsilon_r, \mu_r)]^2 \equiv [\sigma_l^{(1)}(x_0 = 1/4,\ x, , \epsilon_r =$

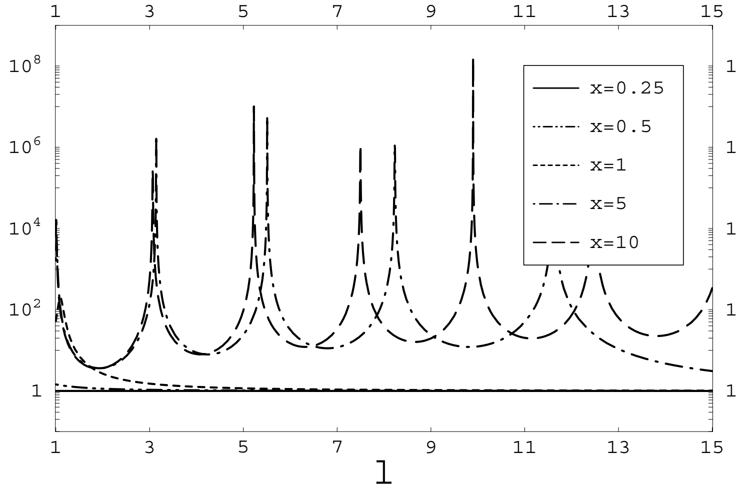


FIG. 4.2. *Normalized singular values* $[\varrho_l^{(1)}]^2$ *versus* l *for* $x_0 = 1/4$ *(quarter-wave case),* $\epsilon_r = +1$, *and a few representative values of* x.

$1)]^2 \,/\, [\sigma_l^{(1)}(x_0 = 1/4 = x, \, , \, \epsilon_r = 1 = \mu_r)]^2$ versus l, parameterized by the normalized wavenumber in the substrate x. From now on, the normalized singular values $[\varrho_l^{(j)}(x_0, x, \epsilon_r, \mu_r)]^2$ defined as

$$\left[\varrho_l^{(j)}(x_0, x, \epsilon_r, \mu_r)\right]^2 \equiv \frac{\left[\sigma_l^{(j)}(x_0, x, \epsilon_r, \mu_r)\right]^2}{\left[\sigma_l^{(j)}(x_0 = x, , \epsilon_r = 1 = \mu_r)\right]^2} \tag{4.1}$$

will be referred to simply as singular values, unless otherwise specified. The singular value spectrum plots for the larger x values considered ($x = 5$ and 10) reveal well-defined resonances (local peaks). The dominant resonances for these larger x values occur around $l \sim \pi$. In fact, the resonances in question appear to arise only when $x \gtrsim 1$. Overall, it is seen that as the material becomes electromagnetically denser, i.e., as the substrate normalized wavenumber x increases, the magnitudes of the singular values become accordingly larger. Since electrically small antennas such as the one considered here can effectively radiate only the lowest multipole orders (such as the dipolar mode), then of particular interest for small antenna applications is the antenna substrate-induced enhancement for low multipolarity l. The plots reveal that the dipolar-mode ($l = 1$) singular values can be significantly higher for the embedding substrate case than for the free-space case. The improvement for $x = 5$ and 10 relative to the free-space case is of more than 3 orders of magnitude (decades). This means that the magnitude of the exciting current or source required for launching of the given dipolar field can be made correspondingly smaller than in free space by embedding the antenna in a high wavenumber or electromagnetically dense substrate. Alternatively, for fixed source energy, the antenna size parameter a can be reduced relative to its value without the embedding substrate. The improvement for $l = 2$ and 3 associated to the larger wavenumber cases ($x = 5$ and 10) is also noticeable.

The respective plot for the case of a resonant or electrically large $x_0 = 10$ antenna is shown in Figure 4.3. The respective magnetic singular value spectra are shown in Figures 4.4–4.5. Many of the key features outlined above in the explanation of the

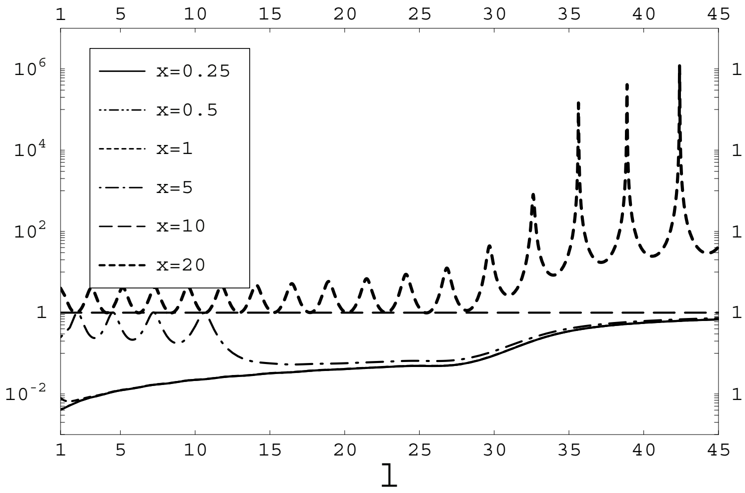


FIG. 4.3. *Normalized singular values* $[\varrho_l^{(1)}]^2$ *versus* l *for* $x_0 = 10$ *(resonant or electrically large antenna),* $\epsilon_r = +1$*, and a few representative values of* x*.*

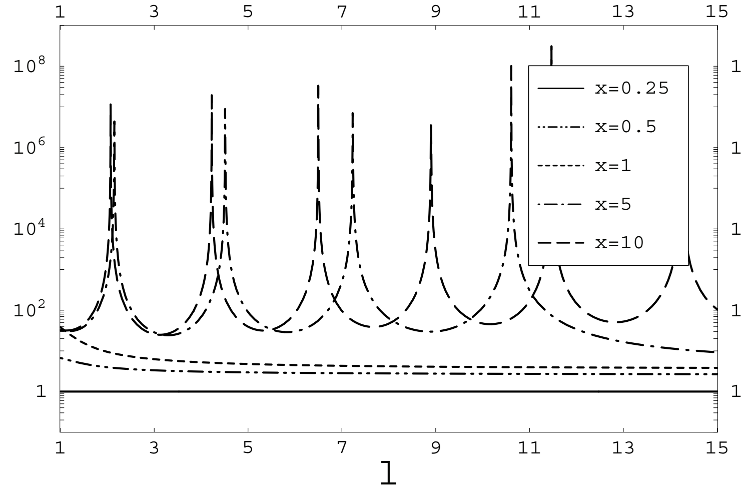


FIG. 4.4. *Normalized singular values* $[\varrho_l^{(2)}]^2$ *versus* l *for* $x_0 = 1/4$ *(quarter-wave case),* $\epsilon_r = +1$*, and a few representative values of* x*.*

particular electric quarter-wave antenna case also arise for these other cases. Yet other aspects become salient. A summary of the main results is given next, along with some of the former observations, as general conclusions learned from these simulations as a whole.

It is seen that, for sufficiently large multipolarity l (i.e., for $l \gtrsim 6$), and for the values of x_0 considered which comprise both small and large or resonant antennas, the singular values are consistently higher for the denser substrates (larger x) than for the less dense substrates including the free-space ($x = x_0$) case. This is true for both electric ($j = 1$) and magnetic ($j = 2$) modes. As we had indicated for the particular electric quarter-wave antenna case, generally for $x = x_0$ (no embedding medium or free-space case), the singular value spectrum decays exponentially with

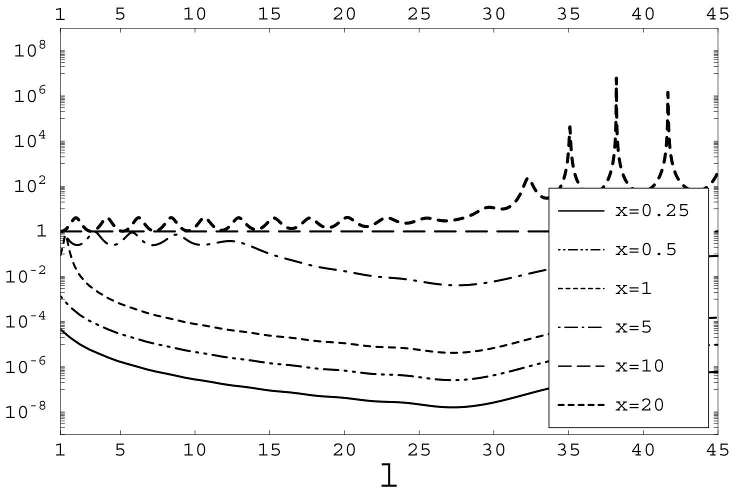


FIG. 4.5. *Normalized singular values $[\varrho_l^{(2)}]^2$ versus l for $x_0 = 10$ (resonant or electrically large antenna), $\epsilon_r = +1$, and a few representative values of x.*

l, i.e., without resonances. This decay is more or less exponential for the smaller antenna cases. For larger antennas the singular values remain more or less within a given order of magnitude until about the cutoff $l \sim k_0 a$ (this value approximately corresponds to the inflection point in the singular value spectrum curve for the free-space case), but this cutoff is clearly higher (includes higher order multipoles) for the large wavenumber cases. (Note that in order to see this visually one would need to plot not the normalized singular values $[\varrho_l^{(j)}]^2$ but the singular values $[\sigma_l^{(j)}]^2$ themselves. Yet a careful comparison of normalized values in Figures 4.2–4.5 and the illustrative free-space values in Figure 4.1 leads to the same conclusion.) This further shows performance enhancement via larger wavenumber or electromagnetically denser substrates since higher multipoles represent higher antenna directivity (higher level of details or narrower width in the radiation pattern). It is also important to note that the enhancement in the singular values due to larger substrate wavenumber k holds for both small and large multipolarities l.

Having shown some of the radiation enhancing possibilities offered by electromagnetically denser substrates, we discuss next the question of local optimal selection of the wavenumber x. Consider, for example, a half-wave antenna (so that $x_0 = 1/2$) embedded in a substrate with $\epsilon_r = 1$ and launching purely magnetic modes ($j = 2$). Local maxima of the respective normalized singular values $[\varrho_l^{(2)}(x_0 = 1/2,\, x,\, \epsilon_r = 1,\, \mu_r)]^2$ for $l = 1$, 2, and 3 were found to occur as follows: for the emission of dipole radiation ($l = 1$) at $x \simeq 1.430$, with an enhancement or gain $[\varrho_l^{(2)}(x_0 = 1/2,\, x = 1.430,\, \epsilon_r = 1,\, \mu_r)]^2 \simeq 3.110 \times 10^5$, relative to free space; for the emission of quadrupole radiation ($l = 2$) at $x \simeq 1.833$, with a gain relative to free space of 1.925×10^7; and for the emission of octupole radiation ($l = 3$) at $x \simeq 2.224$, with a gain of 10^{10}. For antennas embedded in denser substrates the numerical study indicates, however, that the improvement attained is comparatively marginal. For example, the gain associated to going from the aforementioned values of x to the local maxima at $x \sim 10$ is only 44.03, 23.86, and 5.55 for the dipole, quadrupole, and octupole radiation cases, respectively. Conversely, a half-wave antenna radiating purely electric modes instead displays a significantly different behavior in this regard, and the overall improvements of the substrate are also more significant. Thus for modest values of x, a locally maximum improvement in the radiation ability of the half-wave antenna can be attained for the following values: for electric dipole radiation at $x \simeq 0.946$ with a gain $[\varrho_1^{(1)}(x_0 = 1/2,\, x = 0.946,\, \epsilon_r,\, \mu_r)]^2 \simeq 5.36$; for quadrupole radiation at $x \simeq 1.362$ with a gain relative to free space of 163.6; for octupole radiation at $x \simeq 1.800$ with a gain of 2.952×10^4. For denser materials the enhancement relative to free space can be significantly larger. Thus numerical maximization of $[\varrho_l^{(1)}(x_0, x, \epsilon_r, \mu_r)]^2$ yields the following gains associated to going from the aforementioned values of x to the local maxima at $x \sim 10$: 126, 70.81, and 33.56 for the dipole, quadrupole, and octupole radiation cases, respectively. The first two of those numbers are relatively significant enhancements, yet for much denser materials the enhancements are less dramatic, though still meaningful.

A legitimate question arises as to the physical reason behind the appearance of these resonances in the spectra of the non–free-space singular values. As noted earlier, a careful examination of the quantities $F_l^{(j)}$ defined in (2.20) shows that these quantities are essentially the Mie amplitudes, $F_l^{(1)}$ being the amplitudes of the electric modes and $F_l^{(2)}$ those of the magnetic modes [35, 38]. The question that arises now is, Are those resonant peaks, which correspond to local maximum enhancement, related to Mie resonances? Before answering this question we review very briefly the features of Mie resonance that are most relevant to our results. Mie resonances are characterized by the vanishing of the denominators of the amplitudes $F_l^{(j)}$, or, more realistically, by the requirement that those denominators be minimum [38]. Thus the resonance conditions can be cast in the form of approximate transcendental equations, viz.,

$$\sqrt{\epsilon_r}\frac{V_l(x_0)}{h_l^{(+)}(x_0)} \simeq \sqrt{\mu_r}\frac{U_l(x)}{j_l(x)} \tag{4.2}$$

for the electric modes, and

$$\sqrt{\mu_r}\frac{V_l(x_0)}{h_l^{(+)}(x_0)} \simeq \sqrt{\epsilon_r}\frac{U_l(x)}{j_l(x)} \tag{4.3}$$

for the magnetic modes, where U_l and V_l are the functions defined in (2.21) and (2.22). Because of the presence of Bessel functions, (4.2) and (4.3) admit a discrete, albeit infinite, set of solutions. These solutions correspond to the so-called Mie resonances.

Now we can go back to the question of how the observed resonant peaks which correspond to local maximum enhancement relate to Mie resonances. Singular values $[\varrho_l^{(j)}(x_0, x, \epsilon_r, \mu_r)]^2$, defined by (4.1), (3.7), and (3.8), are composed not only of the quantities $|F_l^{(j)}|^2$, defined in (2.20), but also of another term, viz., $[\kappa_l^{(j)}(x_0, x)]^2$, defined in (3.8), and unless these latter quantities are sufficiently well behaved, one cannot conclude anything as to the relationship of the resonant values of $[\varrho_l^{(j)}(x_0, x, \epsilon_r, \mu_r)]^2$ to Mie resonances. Incidentally, the quantities $[\kappa_l^{(j)}(x_0, x)]^2$, where $x \in \mathbb{R}$, are essentially nonpathological combinations of the spherical Bessel functions $j_l(\lambda a)$ which are well behaved for all integer values of l and $\lambda \in \mathbb{R}$ [1] (which represent the most general cases considered in this work). Hence, one can confidently claim that the observed peaks in the spectrum of the singular values are primarily due to the phenomenon of Mie resonance and maximum enhancement conditions are effectively summarized by the two conditions (4.2) and (4.3). Therefore, for a given antenna radiating at a prescribed frequency, the discrete set of solutions x corresponds to a

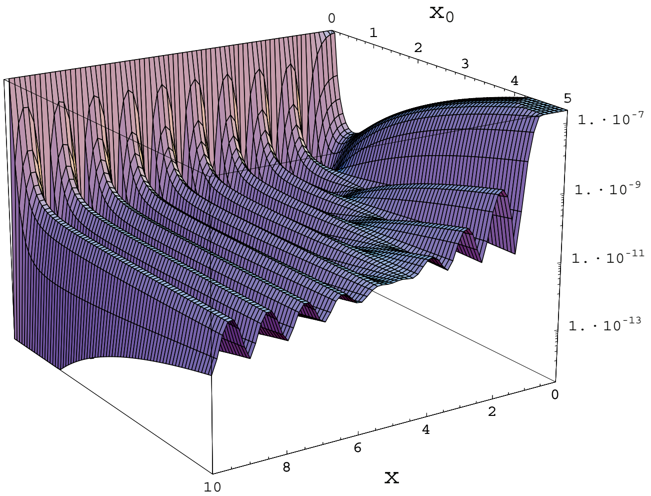


FIG. 4.6. *Logarithmic mesh plot of the source energy $\mathcal{E}_{ME}^{(j=1=l)}$ versus x_0 and x for a double-positive material with $\epsilon_r = +1$.*

set of constitutive parameters ϵ_s and μ_s that maximize the radiated electromagnetic fields. As their amplitudes increase, these radiated fields draw energy from the embedding medium. But because this medium is of finite extent, the energy extraction process saturates, ultimately, and as a result of this saturation the fields fall short of effectively "blowing up."

4.1.2. Further details: Electric dipole radiation. This part examines in greater detail the fundamental electric dipole radiation case, in particular, the multipole moment $a_{l,m}^{(j)} = 1$ if $j = 1 = l$ and $m = 0$, and $a_{l,m}^{(j)} = 0$ otherwise. The minimum source energy reduces in this case to $\mathcal{E}_{ME}^{(j=1=l)}(x_0, x, \epsilon_r, \mu_r) = [\sigma_1^{(1)}(x_0, x, \epsilon_r, \mu_r)]^{-2}$. Figure 4.6 shows a mesh plot of the minimum source energy $\mathcal{E}_{ME}^{(j=1=l)}$ versus the normalized wavenumbers x_0 and x for a double-positive substrate material with $\epsilon_r = 1$. For a double-negative substrate material having $\epsilon_r = -1$, the numerical study shows (plots not shown) that the minimum source energy displays a very similar, though not completely symmetrical, behavior when x changes sign, for a given x_0. Consequently, source energy $\mathcal{E}_{ME}^{(j=1=l)}$ is not an even function of x, and hence distinguishes between double-positive and double-negative embedding substrates. Figure 4.7 shows slices or cross-sections of the mesh plot in Figure 4.6 for particular values of the free-space normalized wavenumber x_0. Similar plots (results not shown) were obtained for $\epsilon_r = -1$ and negative x. Figure 4.6 also shows that, in general terms, source energy tends to decrease as the size of the antenna increases; this is also true when x is negative. Thus as the antenna size increases it tends to be easier to distribute the source currents in a more efficient way. As shown in Figure 4.7, for small antennas the source energy exhibits its first local minima at $|x| \sim 1$. In particular, for $x_0 = 1/4$ (quarter-wave antenna case) and $x_0 = 1/2$ (half-wave case) the first local minimum of $\mathcal{E}_{ME}^{(j=1=l)}$ appears for positive x at $x \simeq 0.960$ and $x \simeq 0.946$, respectively, and for negative x at $x = -0.760$ and $x = -0.860$, respectively. For $x_0 = 1$ (full-wave antenna case) the first local minimum of $\mathcal{E}_{ME}^{(j=1=l)}$ appears for positive x at $x \simeq 1.155$ and for negative x at $x \simeq -1.200$. However, for large antennas a slightly more subtle behavior is observed.

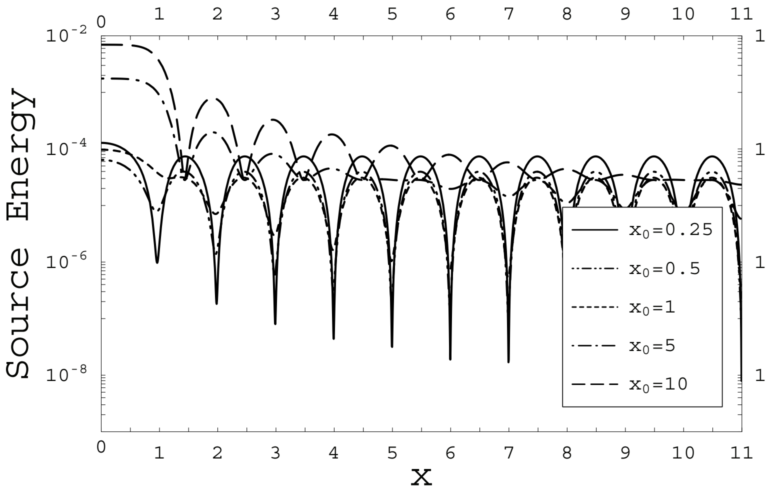


FIG. 4.7. *Logarithmic plot of the source energy $\mathcal{E}_{ME}^{(j=1=l)}$ versus x for $\epsilon_r = +1$ and some representative values of x_0 for a double-positive medium.*

If $|x| < x_0$, the local minima of $\mathcal{E}_{ME}^{(j=1=l)}$ appear at $|x| \sim (2n+1)/2$, $n = 1, 2, 3, \dots$, while if $|x| > x_0$, the minima appear at $|x| \sim n$, $n = 1, 2, 3, \dots$, with the least minimum still belonging to the smallest antenna (cf. Figure 4.7). These rules of thumb depend on the particular combination of constitutive parameters ϵ_r and μ_r under investigation. For example, it was found that for a double-positive material with $\mu_r = 1$ the rules are interchanged; i.e., now, if $x < x_0$, the local minima of $\mathcal{E}_{ME}^{(j=1=l)}$ appear at $x \sim n$, $n = 1, 2, 3, \dots$, while if $x > x_0$, the minima appear at $x \sim (2n+1)/2$, $n = 1, 2, 3, \dots$.

Finally, to further illustrate the possibility of reducing radiator size while achieving a given radiation pattern with prescribed source resources, specifically, source energy, we considered the free-space wavenumber $k_0 = \pi/4$ and sought values of the size parameter a for which the minimum source energy of a source embedded in a medium having $k = 10\pi$ renders the same source energy as a unit-valued a embedded in free space, for which $k = k_0 = \pi/4$. For an embedding substrate with $\epsilon_r = 1$ the first such values of a are $0.098, 0.101, 0.196, 0.204, \dots$ (units of meter), which are seen to occur in pairs around 0.1, 0.2, 0.3, etc. This is not surprising in light of the formula introduced earlier; in particular, the locally optimal values of ka are $ka \sim n\pi$, $n = 1, 2, 3, \dots$, that is, $a \sim 1/10, 2/10, 3/10, \dots$. The values of the size parameter a for which the source energy in question coincides with the free-space case source energy for a larger source having unit-valued radius then occur in pairs around these optimal values, which completes the picture.

4.2. Tuned minimum energy sources: Additional zero reactive power constraint. Next we consider minimum energy sources subjected to the additional zero reactive power constraint. In particular, we require the reactive power to vanish, that is, $\Im[\mathcal{P}] = 0$. As in the preceding subsection, the focus is the fundamental case of an electric dipole radiator (specifically, $a_{l,m}^{(j)} = 1$ if $j = 1 = l$ and $m = 0$, and $a_{l,m}^{(j)} = 0$ otherwise). Particular attention is given to the quarter-wave and the half-wave antenna cases, though some results related to larger antennas are also presented.

As in [32], we define the normalized reactive power

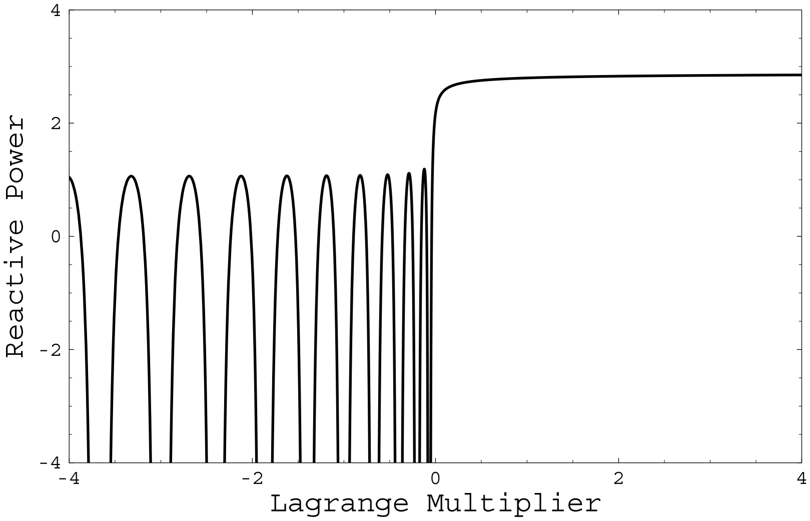


FIG. 4.8. *Plot of the normalized reactive power* $\overline{g}_1^{(1)}$ *versus* χ *for* $x_0 = 1/4 = x$ *and* $\epsilon_r = +1$.

$$\overline{g}_1^{(1)} \equiv \frac{g_1^{(1)}}{\Re\left[\mathcal{P}\right]} = \eta_0 g_1^{(1)}, \tag{4.4}$$

where $\Re\left[\mathcal{P}\right] = 1/2\eta_0$ is the radiated power and where the free-space wave impedance $\eta_0 = \sqrt{\mu_0/\epsilon_0} \simeq 120\pi\,\Omega$. We define χ_0 as the Lagrange multiplier value χ which annuls the normalized reactive power $\overline{g}_1^{(1)}$, i.e., $\overline{g}_1^{(1)}(\chi)\,|_{\chi=\chi_0} = 0$, and for which the resulting source energy is minimal among all such zero reactive power Lagrange multiplier values. The value in question was consistently found to occur in the vicinity of $\chi = 0$. This is not surprising since the absolute or unconstrained minimum energy source and its energy $\mathcal{E}_{ME}^{(j=1=l)}$ correspond to $\chi = 0$; that is, the minimum energy source is $\min \mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)} \equiv \lim_{\chi\to 0} \mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}(x_0, x, \epsilon_r, \mu_r, \chi) = \mathcal{E}_{ME}^{(j=1=l)}$ (see section 3).

Figure 4.8 is a plot of the normalized reactive power $\overline{g}_1^{(1)}$ versus the Lagrange multiplier χ for a quarter-wavelength antenna, embedded in substrates with $\epsilon_r = 1$ and $x = 1/4$. The Lagrange multiplier value χ_0 is sought for which the respective source energy (shown in Figure 4.9) is minimized among all χ values rendering zero reactive power. Tables 4.1, 4.2, 4.3, and 4.4 summarize the values of χ_0, source energy for $\chi = \chi_0$ (i.e., $\mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}$), and absolute minimum energy $\mathcal{E}_{ME}^{(j=1=l)}$ for the case addressed in these plots, as well as for other cases.

One notes, from these results and other plots not shown due to space constraints, that the minimum energy solution $\mathbf{J}_{ME}^{(j=1=l)}$ yields minimum source energy $\mathcal{E}_{ME}^{(j=1=l)}$ or current level, but its reactive power is comparable to the maximum, saturated value corresponding to $\chi \gg 1$ for the double-positive materials and to $\chi \ll -1$ for double-negative materials (cf. Figure 4.8). On the other hand, the new solution $\mathbf{J}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}$ corresponding to χ_0 yields zero reactive power at the expense of a raised source energy or current level (cf. Tables 4.1, 4.2, 4.3, and 4.4). The difference between the source energies $\mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}$ and $\mathcal{E}_{ME}^{(j=1=l)}$ of the two sources $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ and $\mathbf{J}_{ME}$, respectively, is the source energy of the additional nonradiating part contained in $\mathbf{J}_{\mathcal{E},\mathcal{P}}$ whose role in the new source is to counteract the reactive power of the minimum energy source alone. It decreases as the electromagnetic density of the substrate in-

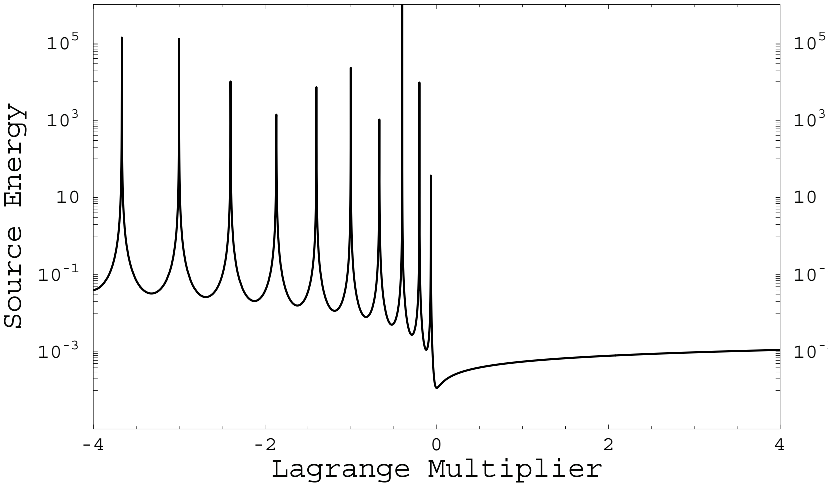


FIG. 4.9. *Plot of the source energy $\mathcal{E}_{\mathcal{EP}}^{(j=1=l)}$ versus χ for $x_0 = 1/4 = x$ and $\epsilon_r = +1$.*

TABLE 4.1
Results of the numerical study for the constrained quarter-wave antenna embedded in a double-positive material with $\epsilon_r = +1$. (The unit of the source energies is A^2/m.)

x	$\chi_0[10^{-4}]$	$\mathcal{E}_{\mathcal{EP}}^{(j=1=l)}(\chi_0)$	$\min \mathcal{E}_{\mathcal{EP}}^{(j=1=l)}$
1/4	-421.2	4.637×10^{-4}	1.161×10^{-4}
1/2	-5.803	8.031×10^{-5}	8.014×10^{-5}
0.511	0	7.809×10^{-5}	7.809×10^{-5}
1	-44.52	2.090×10^{-5}	2.206×10^{-6}
5	-8.039	4.782×10^{-6}	7.094×10^{-8}
10	-4.008	2.486×10^{-6}	1.763×10^{-8}

TABLE 4.2
Results of the numerical study for the constrained half-wave antenna embedded in a double-positive material with $\epsilon_r = +1$. (The unit of the source energies is A^2/m.)

x	$\chi_0[10^{-4}]$	$\mathcal{E}_{\mathcal{EP}}^{(j=1=l)}(\chi_0)$	$\min \mathcal{E}_{\mathcal{EP}}^{(j=1=l)}$
1/4	-899.5	3.087×10^{-4}	5.854×10^{-5}
1/2	-80.37	4.857×10^{-5}	4.191×10^{-5}
1	-35.08	1.513×10^{-5}	8.824×10^{-6}
5	-13.51	3.690×10^{-6}	2.838×10^{-7}
10	-7.234	2.081×10^{-6}	7.051×10^{-8}

creases, this being true for both double-positive and double-negative substrates. We found that, for $x_0 = 1/4$ and $1/2$, performances better than those of the free-space cases (i.e., for which $k = k_0$ and $\epsilon_r = 1 = \mu_r$) can be achieved (cf. Tables 4.1, 4.2, 4.3, and 4.4, though for the sake of space the energy difference $\mathcal{E}_{\mathcal{EP}}^{(j=1=l)}(\chi_0) - \min \mathcal{E}_{\mathcal{EP}}^{(j=1=l)}$ is not explicitly displayed in the tables). Superior performance can also be obtained by means of a judicious choice of the substrate constitutive properties, as we explain below. In addition to this, we note that the minimum of the energy decreases as the electromagnetic density of the substrate increases, whether the substrate is double-positive or double-negative. In Figure 4.9 it is clear that as $\chi \to 0$ the source energy

Table 4.3

Results of the numerical study for the constrained quarter-wave antenna embedded in a double-negative metamaterial with $\epsilon_r = -1$. (The unit of the source energies is A^2/m).

x	$\chi_0[10^{-4}]$	$\mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}(\chi_0)$	$\min \mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}$
$-1/4$	$\cdot$	$\cdot$	3.071×10^{-5}
$-1/2$	$\cdot$	$\cdot$	1.462×10^{-5}
-1.338	0	7.980×10^{-5}	7.980×10^{-5}
-1	24.27	6.288×10^{-5}	2.540×10^{-5}
-5	7.218	1.708×10^{-5}	8.167×10^{-7}
-10	3.757	9.392×10^{-6}	2.029×10^{-7}

Table 4.4

Results of the numerical study for the constrained half-wave antenna embedded in a double-negative metamaterial with $\epsilon_r = -1$. (The unit of the source energies is A^2/m.)

x	$\chi_0[10^{-4}]$	$\mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}(\chi_0)$	$\min \mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}$
$-1/4$	$\cdot$	$\cdot$	3.718×10^{-5}
$-1/2$	$\cdot$	$\cdot$	2.554×10^{-5}
-1	37.88	2.696×10^{-5}	1.462×10^{-5}
-5	13.74	6.821×10^{-6}	4.702×10^{-7}
-10	7.304	3.825×10^{-6}	1.168×10^{-7}

$\mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}$ reaches an absolute minimum $\min \mathcal{E}_{\mathcal{E}\mathcal{P}}^{(j=1=l)}$, as expected. This minimum is not the same for double-positive materials and double-negative metamaterials (cf. Tables 4.1, 4.2, 4.3, and 4.4), as we discussed earlier. Interestingly, the cancellation of the reactive power is not always possible. For instance, for a quarter-wave antenna and for $\epsilon_r = 1$, the equation $\overline{g}_1^{(1)}(\chi)\,|_{\chi=\chi_0} = 0$ admits no solutions if $x = -1/4$ or $-1/2$. (This is also true for $x_0 = 1/2$ and 1.)

Furthermore, it also follows that, if one allows the electromagnetic properties of the embedding substrate (i.e., ϵ_r and μ_r) to vary, then one could make the reactive power vanish for $\chi_0 = 0$, this being a *matching condition* under which the minimum energy sources are not only of local minimum energy (see below) but also self-matched to resonance. Let us illustrate this for a quarter-wavelength antenna. For a given positive relative electric permittivity, for instance, $\epsilon_r = 1$, we find that the matching condition mentioned above is satisfied for $x \simeq 0.511$, i.e., in this case $\epsilon_r = 1$ and $\mu_r \simeq 4.18$. Now, for a given negative relative electric permittivity, for instance, $\epsilon_r = -1$, we find that the matching condition is satisfied for $x \simeq -1.338$, i.e., in this case $\epsilon_r = -1$ and $\mu_r = -28.64$. A word of caution is necessary at this point. From the definition itself of the abovementioned matching, for $\chi_0 = 0$ one obtains $\mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)} = \min \mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}$ (cf. Tables 4.1 and 4.3). Yet one must not be lured into thinking that the substrate constitutive properties ϵ_r and μ_r associated to the matching cases must correspond to global minima for $\mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}$, i.e., that they represent the best substrate values. This is very clearly illustrated in Tables 4.1 and 4.3 where for $x = 1$, 5, and 10 in Table 4.1 and for $x = -1$, -5, and -10 in Table 4.3 one has $\mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}|_{\chi_0 \neq 0} < \mathcal{E}_{\mathcal{E},\mathcal{P}}^{(j=1=l)}|_{\chi_0 = 0}$. In other words a quarter-wavelength antenna embedded in substrates having those values of x as their electromagnetic densities exhibits source energies smaller than those exhibited by the antenna when it is embedded in a substrate whose constitutive parameters satisfy the matching condition.

Appendix A. Wavefunctions $\mathfrak{B}^{(j)}_{l,m}$ for piecewise-constant radially symmetric backgrounds. The aim of this appendix is to show that the multipole moments $a^{(j)}_{l,m}$ are given by (2.17) with the wavefunctions $\mathfrak{B}^{(j)}_{l,m}$ given by (2.19) and (2.20). The manipulations below rely on (2.13) and (2.14) and the concept of reciprocity. The latter can be stated as follows [7]: The reaction (coupling) $\mathcal{R}_{\mathbf{E}\to\mathbf{J}_0}$ of the field $\mathbf{E}$ produced by a source $\mathbf{J}$ on another source $\mathbf{J}_0$, given by

$$\mathcal{R}_{\mathbf{E}\to\mathbf{J}_0} = \int d\mathbf{r}\mathbf{E}(\mathbf{r})\cdot\mathbf{J}_0(\mathbf{r}), \tag{A.1}$$

is equal to the reaction of the field $\mathbf{E}_0$ produced by the source $\mathbf{J}_0$ on the source $\mathbf{J}$, in particular,

$$\mathcal{R}_{\mathbf{E}_0\to\mathbf{J}} = \int d\mathbf{r}\mathbf{E}_0(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}) = \mathcal{R}_{\mathbf{E}\to\mathbf{J}_0}. \tag{A.2}$$

To evaluate the field due to a current distribution $\mathbf{J}(\mathbf{r})$ that is embedded in the piecewise-constant radially symmetric background of interest, we consider, without loss of generality, the following two classes of canonical sources:

$$\left[\mathbf{J}^{(1)}_{l,m}\right]_0(\mathbf{r}) = \delta(r-R)\hat{\mathbf{r}}\times\mathbf{Y}_{l,m}(\hat{\mathbf{r}}) \tag{A.3}$$

and

$$\left[\mathbf{J}^{(2)}_{l,m}\right]_0(\mathbf{r}) = \delta(r-R)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), \tag{A.4}$$

where in both expressions $R > a$ represents the radius of the helper source centered around the origin. (Ultimately, the multipole moments $a^{(j)}_{l,m}$ are expected to, and in fact will, turn out to be independent of R.) The justification for the above considerations relies on two results: (1) the transverse component of an arbitrary vector field on the spherical surface of radius $R > a$ centered about the origin is *uniquely* characterized by its expansion in terms of the vector spherical harmonics $\mathbf{Y}_{l,m}(\hat{\mathbf{r}})$ and their associated vector functions $\hat{\mathbf{r}}\times\mathbf{Y}_{l,m}(\hat{\mathbf{r}})$, and (2) the multipole moments characterizing any electric field outside the support of the emitting sources is *uniquely* determined by the tangential component of this field on a sphere totally enclosing the support of the emitting sources (as noted in section 2).

The field $[\mathbf{E}^{(j)}_{l,m}]_{inc}$ that would be produced by the source $[\mathbf{J}^{(j)}_{l,m}]_0$ in (A.4) if it were in free space (this will be the incident field in the following) is given by

$$\left[\mathbf{E}^{(j)}_{l,m}\right]_{inc} = \int d\mathbf{r}\bar{\mathbf{G}}_0(\mathbf{r},\mathbf{r}')\cdot\left[\mathbf{J}^{(j)}_{l,m}\right]_0,$$

where $\bar{\mathbf{G}}_0(\mathbf{r},\mathbf{r}')$ is the multipole representation of the free-space electric dyadic Green's function [39], viz.,

$$\begin{aligned}\bar{\mathbf{G}}_0(\mathbf{r},\mathbf{r}') = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\frac{-\omega\mu_0}{k_0 l\,(l+1)}&\Big\{k_0^2\left[j_l(k_0r_<)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}_<)\right]\left[h_l^{(+)}(k_0r_>)\mathbf{Y}^*_{l,m}(\hat{\mathbf{r}}_>)\right]\\ &+\boldsymbol{\nabla}\times\left[j_l(k_0r_<)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}_<)\right]\boldsymbol{\nabla}\times\left[h_l^{(+)}(k_0r_>)\mathbf{Y}^*_{l,m}(\hat{\mathbf{r}}_>)\right]\Big\}\\ &+\frac{i}{\omega\epsilon_0}\hat{\mathbf{r}}\hat{\mathbf{r}}\delta\left(\mathbf{r}-\mathbf{r}'\right).\end{aligned} \tag{A.5}$$

The $<$ ($>$) subscript designates the smaller (larger) of r and r'.

For $r < R$ the field $[\mathbf{E}^{(1)}_{l,m}]_{inc}$ is found to be given by

$$\left[\mathbf{E}^{(1)}_{l,m}\right]_{inc}(\mathbf{r}) = \tau_l(k_0, R)\boldsymbol{\nabla} \times [j_l(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], \tag{A.6}$$

where we have defined

$$\tau_l(k_0, R) \equiv -\eta_0 k_0 R^2 V_l\,(k_0 R)\,. \tag{A.7}$$

Along analogous lines, the (incident) field $[\mathbf{E}^{(2)}_{l,m}]_{inc}$ produced by the source $[\mathbf{J}^{(1)}_{l,m}]_0$ in (A.3) in free space is found, for $r < R$, to be given by

$$\left[\mathbf{E}^{(2)}_{l,m}\right]_{inc}(\mathbf{r}) = \zeta_l(k_0 R) j_l(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), \tag{A.8}$$

where we have introduced

$$\zeta_l(k_0 R) \equiv -\eta_0\,(k_0 R)^2\, h_l^{(+)}(k_0 R). \tag{A.9}$$

The obtainment of the above results requires the use of orthogonality properties of the vector spherical harmonics $\mathbf{Y}_{l,m}(\hat{\mathbf{r}})$ and the associated vector functions $\hat{\mathbf{r}} \times \mathbf{Y}_{l,m}(\hat{\mathbf{r}})$.

Introducing the index of refraction $n = \sqrt{\mu_s\epsilon_s}/\sqrt{\mu_0\epsilon_0} = \sqrt{\mu_r\epsilon_r}$, the evaluation of the total fields $[\mathbf{E}^{(j)}_{l,m}]_0$, $j = 1, 2$, associated with these sources for $r < R$ is seen from (1.1), (1.2), (2.1)–(2.3) to correspond to the solution of the forward scattering problem associated for $r < R$ with the equation

$$\left[\boldsymbol{\nabla}^2 + k_0^2\Theta(r-a) + n^2 k_0^2\Theta(a-r)\right]\left[\mathbf{E}^{(j)}_{l,m}\right]_0(\mathbf{r}) = \mathbf{0} \tag{A.10}$$

upon excitation by the incident fields $[\mathbf{E}^{(j)}_{l,m}]_{inc}$.

The total (incident plus scattered) field $[\mathbf{E}^{(1)}_{l,m}]_0$ must be, due to considerations of causality (in the scattered field) and well-behavedness of the interior field for $r < a$, of the form

$$\left[\mathbf{E}^{(1)}_{l,m}\right]_0(\mathbf{r}) = \begin{cases} \boldsymbol{\nabla} \times \left[\tau_l(k_0, R) j_l(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}) + D_1 h_l^{(+)}(k_0 r)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})\right], & r > a, \\ A_1\boldsymbol{\nabla} \times [j_l(kr)\mathbf{Y}_{l,m}(\hat{\mathbf{r}})], \quad r \le a, \end{cases} \tag{A.11}$$

where $k = nk_0$ is the wavenumber of the field in the background material confined within the source volume V and A_1 and D_1 are coefficients that are to be determined by imposing continuity of the tangential components of the electric and magnetic fields on the boundary $\partial V \equiv \{\mathbf{r} \in \mathbb{R}^3 : r = a\}$. Analogously, the total field $[\mathbf{E}^{(2)}_{l,m}]_0$ must be of the form

$$\left[\mathbf{E}^{(2)}_{l,m}\right]_0(\mathbf{r}) = \begin{cases} \left[\zeta_l(k_0 R) j_l(k_0 r) + D_2 h_l^{(+)}(k_0 r)\right] \times \mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & r > a, \\ A_2 j_l(kr)\mathbf{Y}_{l,m}(\hat{\mathbf{r}}), & r \le a, \end{cases} \tag{A.12}$$

where A_2 and D_2 are coefficients that need to be determined from the boundary conditions on ∂V.

By imposing the continuity requirements on the boundary ∂V and using the Wronskian relation for spherical Bessel functions [6], one obtains (for $j = 1$)

$$\begin{aligned}\frac{A_1}{\tau_l(k_0, R)} &= \frac{i/\,(k_0 a)\,(ka)}{(\epsilon_r/\mu_r)^{1/2}\, j_l(ka) V_l(k_0 a) - h_l^{(+)}(k_0 a) U_l(ka)} \\ &\equiv F_l^{(1)}\left(k_0 a, ka, \epsilon_r, \mu_r\right), \qquad \text{(A.13)}\end{aligned}$$

where U_l and V_l have already been defined in (2.21) and (2.22). The coefficient A_2 associated with the field $[\mathbf{E}_{l,m}^{(2)}]_0$ can be obtained by an analogous procedure which yields

$$\begin{aligned}\frac{A_2}{\zeta_l(k_0 R)} &= \frac{i\mu_r/\,(k_0 a)\,(ka)}{(\mu_r/\epsilon_r)^{1/2}\, j_l(ka) V_l(k_0 a) - h_l^{(+)}(k_0 a) U_l(ka)} \\ &\equiv F_l^{(2)}\left(k_0 a, ka, \epsilon_r, \mu_r\right). \qquad \text{(A.14)}\end{aligned}$$

Along with (A.11) and (A.12) the above results define the fields $[\mathbf{E}_{l,m}^{(1)}]_0$ and $[\mathbf{E}_{l,m}^{(2)}]_0$ in the region V.

By applying the reciprocity relation equation (A.2) to the preceding results (in particular, (A.3), (A.4), (A.7), (A.9), (A.11)–(A.14)), one finds that the multipole moments $a_{l,m}^{(j)}$ are indeed independent of R and are given by (2.17) with the source-free wavefunctions $\mathfrak{B}_{l,m}^{(j)}(\mathbf{r})$ given by (2.19)–(2.22). In obtaining these results we have also recalled the multipole expansion for the electric field $\mathbf{E}(\mathbf{r})$, i.e., (2.13) and (2.14), along with the orthogonality and analytic continuation properties of the vector spherical harmonics, and the analytic continuation property of the spherical Bessel functions of the first kind, viz., $j_l^*(ka) = j_l(k^* a)$.

Acknowledgments. E.A.M. gratefully acknowledges enlightening conversations with Prof. Anthony Devaney, Dr. Arthur Yaghjian, and Dr. Arje Nachman that were pertinent to the research reported here. Also, the authors would like to express their gratitude to the referees for their interesting comments and suggestions.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] E. E. Altshuler, *Electrically small genetic antennas immersed in a dielectric*, in Proceedings of the IEEE Antennas and Propagation Society International Symposium, Vol. 3, 2004, pp. 2317–2320.

[3] A. Alú and N. Engheta, *Radiation from a traveling-wave current sheet at the interface between a conventional material and a metamaterial with negative permittivity and permeability*, Microw. Opt. Technol. Lett., 35 (2002), pp. 460–463.

[4] T. S. Angell and A. Kirsch, *Optimization Methods in Electromagnetic Radiation*, Springer-Verlag, New York, 2004.

[5] T. S. Angell, A. Kirsch, and R. E. Kleinman, *Antenna control and optimization*, Proc. IEEE, 79 (1991), pp. 1559–1568.

[6] G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, Academic Press, New York, 2005.

[7] C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.

[8] M. S. Berger, *Nonlinear Functional Analysis*, Academic Press, New York, 1977.

[9] O. M. Bucci, G. D'Elia, G. Mazzarella, and G. Panariello, *Antenna pattern synthesis: A new general approach*, Proc. IEEE, 82 (1994), pp. 358–371.

[10] K. Buell, H. Mosallaei, and K. Sarabandi, *A substrate for small patch antennas providing tunable miniaturization factors*, IEEE Trans. Microw. Theory Tech., 54 (2006), pp. 135–146.

[11] R. E. Collin, *Field Theory of Guided Waves*, IEEE Press, New York, 1991.

[12] A. J. Devaney, E. A. Marengo, and M. Li, *Inverse source problem in nonhomogeneous background media*, SIAM J. Appl. Math., 67 (2007), pp. 1353–1378.

[13] A. J. Devaney and R. P. Porter, *Holography and the inverse source problem. Part* II: *Inhomogeneous media*, J. Opt. Soc. Amer. A, 2 (1985), pp. 2006–2011.

[14] A. J. Devaney and E. Wolf, *Multipole expansions and plane wave representations of the electromagnetic field*, J. Math. Phys., 15 (1974), pp. 234–244.

[15] N. Engheta, A. Alù, R. W. Ziolkowski, and A. Erentok, *Fundamentals of waveguide and antenna applications involving DNG and SNG metamaterials*, in Metamaterials: Physics and Engineering Explorations, N. Engheta and R. W. Ziolkowski, eds., Wiley-IEEE Press, New York, 2006, pp. 43–85.

[16] S. Enoch, G. Tayeb, P. Sabouroux, N. Guérin, and P. Vincent, *A metamaterial for directive emission*, Phys. Rev. Lett., 89 (2002), 213902.

[17] A. Erentok, P. L. Luljak, and R. W. Ziolkowski, *Characterization of a volumetric metamaterial realization of an artificial magnetic conductor for antenna applications*, IEEE Trans. Antennas and Propagation, 53 (2005), pp. 160–172.

[18] R. C. Hansen and M. Burke, *Antennas with magneto-dielectrics*, Microw. Opt. Technol. Lett., 26 (2000), pp. 75–78.

[19] R. F. Harrington, *On the gain and beamwidth of directional antennas*, IRE Trans. Antennas and Propagation, 6 (1958), pp. 219–225.

[20] E. L. Hill, *The theory of vector spherical harmonics*, Amer. J. Phys., 22 (1954), pp. 211–214.

[21] J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1999.

[22] J. R. James and J. C. Vardaxoglou, *Investigation of properties of electrically-small spherical ceramic antennas*, Electron. Lett., 38 (2002), pp. 1160–1162.

[23] A. J. Kurdilla and M. Zabarankin, *Convex Functional Analysis*, Birkhäuser-Verlag, Basel, 2005.

[24] D. Lamensdorf, *An experimental investigation of dielectric-coated antennas*, IEEE Trans. Antennas and Propagation, 15 (1967), pp. 767–771.

[25] D. Lamensdorf and C.-Y. Ting, *An experimental and theoretical study of the monopole embedded in a cylinder of anisotropic dielectric*, IEEE Trans. Antennas and Propagation, 16 (1968), pp. 342–349.

[26] K. W. Leung, *Complex resonance and radiation of hemispherical dielectric-resonator antenna with a concentric conductor*, IEEE Trans. Microw. Theory Tech., 49 (2001), pp. 524–531.

[27] S. A. Long, M. W. McAllister, and L. C. Shen, *The resonant cylindrical dielectric cavity antenna*, IEEE Trans. Antennas and Propagation, 31 (1983), pp. 406–412.

[28] G. Lovat, P. P. Burghignoli, F. Capolino, D. R. Jackson, and D. R. Wilton, *Analysis of directive radiation from a line source in a metamaterial slab with low permittivity*, IEEE Trans. Antennas and Propagation, 54 (2006), pp. 1017–1030.

[29] Y. Mano and S. Bae, *A small meander antenna by magneto-dielectric material*, in Proceedings of the IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, Vol. 1, 2005, pp. 63–66.

[30] E. A. Marengo and A. J. Devaney, *Time-dependent plane wave and multipole expansions of the electromagnetic field*, J. Math. Phys., 39 (1998), pp. 3643–3660.

[31] E. A. Marengo and A. J. Devaney, *The inverse source problem of electromagnetics: Linear inversion formulation and minimum energy solution*, IEEE Trans. Antennas and Propagation, 47 (1999), pp. 410–412.

[32] E. A. Marengo, A. J. Devaney, and F. K. Gruber, *Inverse source problem with reactive power constraint*, IEEE Trans. Antennas and Propagation, 52 (2004), pp. 1586–1595.

[33] E. A. Marengo and R. W. Ziolkowski, *A new procedure for specifying nonradiating current distributions and the fields they produce*, J. Math. Phys., 41 (2000), pp. 845–866.

[34] E. A. Marengo and R. W. Ziolkowski, *Nonradiating and minimum energy sources and their fields: Generalized source inversion theory and applications*, IEEE Trans. Antennas and Propagation, 48 (2000), pp. 1553–1562.

[35] G. Mie, *Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen*, Ann. Phys. (8), 25 (1908), pp. 377–445.

[36] H. Mosallaei and K. Sarabandi, *Magneto-dielectrics in electromagnetics: Concept and applications*, IEEE Trans. Antennas and Propagation, 52 (2004), pp. 1558–1567.

[37] H. R. Raemer, *Radiation from linear electric or magnetic antennas surrounded by a spherical plasma shell*, IRE Trans. Antennas and Propagation, 10 (1962), pp. 69–78.

[38] J. A. Stratton, *Electromagnetic Theory*, McGraw–Hill, New York, 1941.

[39] C. T. Tai, *Dyadic Green's Functions in Electromagnetic Theory*, Intext Educational, Scranton, PA, 1971.

[40] B. Temelkuran, M. Bayindir, E. Ozbay, R. Biswas, M. M. Sigalas, G. Tuttle, and K. M. Ho, *Photonic crystal-based resonant antenna with a very high directivity*, J. Appl. Phys., 87 (2000), pp. 603–605.

[41] M. Thèvenot, C. Cheype, A. Reineix, and B. Jecko, *Directive photonic-bandgap antennas*, IEEE Trans. Microw. Theory Tech., 47 (1999), pp. 2115–2122.

[42] D. Tonn and R. Bansal, *Practical considerations for increasing radiated power from an electrically small antenna by application of a double negative metamaterial*, in Proceedings of the IEEE Antennas and Propagation Society International Symposium, Vol. 2A, 2005, pp. 602–605.

[43] L. Tsang, A. Ishimaru, R. P. Porter, and D. Rouseff, *Holography and the inverse source problem.* III. *Inhomogeneous attenuative media*, J. Opt. Soc. Amer. A, 4 (1987), pp. 1783–1787.

[44] H. Tuy and N. V. Thuong, *On the global minimization of a convex function under general nonconvex constraints*, Appl. Math. Optim., 18 (1988), pp. 119–142.

[45] H. A. Wheeler, *Fundamental limitations of small antennas*, Proc. IRE, 35 (1947), pp. 1479–1484.

[46] H. A. Wheeler, *Small antennas*, IEEE Trans. Antennas and Propagation, AP-23 (1975), pp. 462–469.

[47] B.-I. Wu, W. Wang, J. Pacheco, X. Chen, T. Grzegorczyk, and J. Kong, *A study of using metamaterials as antenna substrate to enhance gain*, Progr. Electromag. Res., 51 (2005), pp. 295–328.

[48] R. W. Ziolkowski and A. D. Kipple, *Application of double negative materials to increase the power radiated by electrically small antennas*, IEEE Trans. Antennas and Propagation, 51 (2003), pp. 2626–2640.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 111–125

BOUNDARY DRIVEN WAVEGUIDE ARRAYS: SUPRATRANSMISSION AND SADDLE-NODE BIFURCATION[*]

HADI SUSANTO[†]

Abstract. In this paper, we consider a semi-infinite discrete nonlinear Schrödinger equation driven at one edge by a driving force. The equation models the dynamics of coupled waveguide arrays. When the frequency of the forcing is in the allowed band of the system, there will be a linear transmission of energy through the lattice. Yet, if the frequency is in the upper forbidden band, then there is a critical driving amplitude for a nonlinear tunneling, which is called supratransmission, of energy to occur. Here, we discuss mathematically the mechanism and the source of the supratransmission. By analyzing the existence and the stability of the rapidly decaying static discrete solitons of the system, we show rigorously that two of the static solitons emerge and disappear in a saddle-node bifurcation at a critical driving amplitude. One of the emerging solitons is always stable in its existence region and the other is always unstable. We argue that the critical amplitude for supratransmission is then the same as the critical driving amplitude of the saddle-node bifurcation. We consider as well the case of the forcing frequency in the lower forbidden band. It is discussed briefly that there is no supratransmission because in this case there is only one rapidly decaying static soliton that exists and is stable for any driving amplitude.

Key words. supratransmission, nonlinear tunneling, saddle-node bifurcations

AMS subject classifications. 34D35, 35Q53, 37K50, 39A11, 78A40

DOI. 10.1137/070698828

1. Introduction.

An exotic nonlinear phenomenon has been discovered recently by Geniet and Leon [1] in a semi-infinite chain of coupled oscillators driven at one edge by a time periodic forcing. Energy excitations will propagate through the chain if the driving frequency is in the allowed band of the discrete system. It is natural because of the system's dispersion relation. In contrast, it would be expected that if the forcing frequency is in the band gap, then there would be no energy flow. Yet, Geniet and Leon [1] show theoretically and experimentally that there is a definite driving amplitude threshold above which a sudden energy flow takes place. This phenomenon is called nonlinear supratransmission [1]. An exciting independent work on a modified Klein–Gordon equation describing the Josephson phase of layered high-$\mathrm{T_c}$ superconductors shows the presence of the same phenomenon [2]. Promising technological applications employing supratransmission have been proposed accompanying these findings, such as binary signal transmissions of information [3] and terahertz frequency selection devices [4].

In [5] Khomeriki considers boundary driven coupled optical focusing waveguide arrays described by

$$i\frac{\partial\psi_n}{\partial z} = -\psi_{n+1} - \psi_{n-1} - \gamma|\psi_n|^2\psi_n, \quad \psi_0 = Ae^{i\Delta z}, \tag{1.1}$$

with $\gamma > 0$ and $n = 1, 2, \ldots$. Here, ψ_n is the electromagnetic wave amplitude in the nth guiding core, z is the propagation variable, Δ is the propagation constant or the driving frequency, and γ represents the nonlinearity coefficient, which is taken

[*] Received by the editors July 30, 2007; accepted for publication (in revised form) May 27, 2008; published electronically August 27, 2008.
http://www.siam.org/journals/siap/69-1/69882.html

[†] School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (hadi.susanto@nottingham.ac.uk).

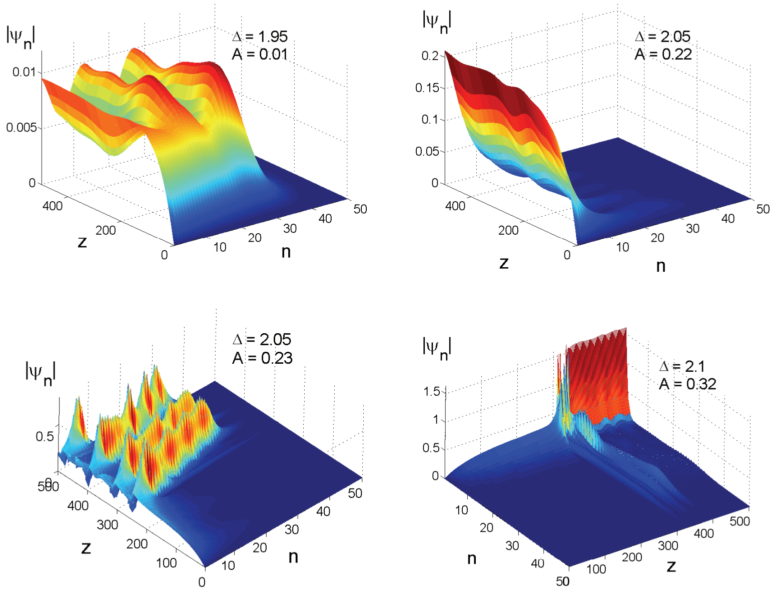


FIG. 1.1. *Three-dimensional plots of time evolution of the boundary driven waveguide arrays of* (1.1). *When the driving frequency $-2 < \Delta < 2$ is in the allowed band, any driving amplitude will lead to an energy flow to remote sites (top left panel). If Δ is in the upper forbidden band and A is small enough, the boundary will excite a couple of arrays only (top right panel). Yet, there is a critical threshold amplitude $A_{\text{th}}(\Delta)$ above which there is a nonlinear forbidden band tunneling indicated by the released of a train of discrete solitons (bottom left panel). A quantitatively different behavior of supratransmission occurs when the driving frequency is large enough, as is shown in the bottom right panel.*

to be $\gamma = 2$ in this report. This model can also be considered as a slow modulation wave approximation to the discrete sine-Gordon equation [6]. Similar to the nonlinear band-gap tunneling observed by Geniet and Leon [1], it is reported that there is a critical threshold $A_{\text{th}}(\Delta)$ for supratransmission when the propagation constant Δ is in the forbidden band $\Delta > 2$ [5].

In Figure 1.1, we present numerical simulations of the dynamics of (1.1). Following [5], the driving is turned on adiabatically to avoid the appearance of an initial shock by assuming the form

$$A = \breve{A}(1 - \exp(-z/\tau)),$$

where we omit the breve symbol henceforth. In the following figures, we take $\tau = 50$ and apply a linearly increasing damping to the last 20 sites to suppress edge reflection.

Presented in the top left panel of Figure 1.1 is a three-dimensional plot of time evolution of (1.1) when the driving frequency is in the allowed band $-2 < \Delta < 2$. A small driving amplitude will excite all the sites. On the other hand, if the driving frequency is in the upper band $\Delta > 2$, a small A will only excite several neighboring sites, as is shown in the top right panel of Figure 1.1. Yet, if the driving amplitude is large enough, then a train of "traveling" discrete solitons can be released [5] (see bottom left panel of the same figure). This flow of energy is the so-called supratransmission or nonlinear forbidden band tunneling, and we call the minimum A for supratransmission to occur a critical threshold A_{th}. The word *traveling* is in quotes because (1.1) does not admit a genuine one (see [7] and the references therein). If one waits long enough, the gap solitons will be trapped by the lattice. Khomeriki [5] also notices an immediate trapping when the driving frequency Δ is relatively large, as is shown in the bottom right panel of Figure 1.1. In this regime, the corresponding discrete gap solitons are highly localized.

An analytical approximation of $A_{\mathrm{th}}(\Delta)$ in the limit $0 < \Delta - 2 \ll 1$ is given by [5]

$$A_{\mathrm{th}}(\Delta) = \sqrt{\Delta - 2}. \tag{1.2}$$

Remark 1.1. Equation (1.1) is symmetric with respect to the transformation $A \to -A$ and $\psi_n \to -\psi_n$. This means that there is also a critical amplitude $-A_{\mathrm{th}}(\Delta)$ if one applies $A < 0$ such that for $A < -A_{\mathrm{th}}(\Delta) < 0$, a nonlinear forbidden band tunneling will occur. Equation (1.1) is also symmetric with respect to the transformation $\Delta \to -\Delta$, $\psi_n \to (-1)^n \psi_n$, and $\gamma \to -\gamma$. Therefore, the same phenomenon can be observed in defocusing waveguide arrays $\gamma < 0$. Due to the transformation, the only difference of defocusing arrays from the self-focusing ones is that there will be a π phase difference between neighboring lattices.

It is presented in [5] that the numerical result for the threshold amplitude deviates rapidly from the approximation (1.2). It is because (1.2) is actually the amplitude-temporal frequency relation of the continuous nonlinear Schrödinger (NLS) equation's solitons. The relation has been phase-shifted properly due to some transformation, i.e., $\psi_n \to \psi_n \exp(2iz)$. Applying the transformation to (1.1) will take it to a normalized standard finite difference approximation of the continuous NLS equation.

Remembering the aforementioned promising applications of nonlinear tunneling, it is therefore of interest to obtain an approximation of the threshold amplitude in the other limit $\Delta - 2 \gg 1$. This is one of the aims of the present report. The other aim is to understand mathematically the mechanism of the nonlinear tunneling. It is mentioned but not rigorously proved [8] that the supratransmission happens because of the emergence of two solutions at the critical driving amplitude, i.e., a saddle-node bifurcation. If this is the case, then this means that supratranmissions correspond to an existence issue, as opposed to a stability phenomenon [9], and the threshold amplitude is not necessarily the amplitude of the corresponding fundamental soliton (1.2). Understanding the source of supratransmission also will allow us to explain, for example, the reason why there is no threshold amplitude for nonlinear tunneling when the driving frequency is in the lower forbidden band $\Delta < -2$.

Nonetheless, one may question the relevance of our first aim, as supratransmission is quickly trapped by the lattices for large Δ. Even though our analysis may not be immediately applicable to the present case, the aim is still of relevance. There are several experimentally realizable discrete equations that support "traveling" solitons in a parameter region where the gap solitons are highly localized. One particular example is the discrete Schrödinger equation with saturable nonlinearity in the large nonlinearity coefficient regime [10]. We have observed supratransmission in this equation and have successfully applied our analysis presented herein to obtain an approximation to the threshold amplitude [11]. Later on in this paper, we also conjecture that our analytically obtained approximation, presented in terms of a power series expansion, may well be convergent uniformly in the region of interest, i.e., $\Delta > 2$. Moreover, the mathematical procedure presented herein can also be applied as an alternative method to analyze the bistability effect considered, e.g., in [12]. We might even consider it simpler and more appropriate as the analysis can then be done solely in its discrete setup, with no need for approximating the problem with its continuous counterpart [12].

In this study, we will show that the supratransmission is indeed related to saddle-node bifurcations. To mathematically prove this, our strategy is as follows. We will first prove the existence of a mode bifurcating from the constant solution $\psi_n \equiv 0$ due to the driving site. We will also show that there is a singular mode bifurcating from infinity. We will then demonstrate that these two modes collide in a saddle-node bifurcation by developing an asymptotic analysis in the range of Δ large. Such an analysis is doable in that regime because the modes are highly localized. The final step to show that the critical amplitude is the same as the threshold amplitude for supratransmission is to prove that the mode bifurcating from the zero state is stable, all the way on its existence region. Using this result, then we can derive an approximation of the threshold amplitude in terms of a power series expansion that can be calculated to any order. Numerical computations will be presented as well to compare our analytical results.

Our paper is outlined as follows. In subsection 2.1, we present our asymptotic analysis for the existence of monotonically decaying static solutions of (1.1). The next subsection will contain our study on the stability analysis of solutions discussed in the preceding subsection. Using the same procedures, we then briefly discuss in subsection 2.3 that there is no supratransmission in the case of $\Delta < -2$ as there is no bifurcation occurring in this regime. Then we compare our analytical findings with the results of numerical computations in subsection 2.4. Finally, we summarize our findings and present our conclusions in the last section.

2. Existence and stability analysis of rapidly decaying discrete solitons.

2.1. Existence analysis. Stationary solutions of (1.1) are sought in the form of $\psi_n(z) = \phi_n e^{i\Delta z}$, where ϕ_n is a real-valued function. This ansatz can be applied, as one would naturally expect that all the sites will be excited with the same frequency as the driving frequency. Since we are interested in the large propagation constant Δ, we scale $\Delta \to 1$ and define $\epsilon = 1/\Delta$. Hence, we consider $|\epsilon| \ll 1$. A static equation of (1.1) is then given by

$$F(\phi, \epsilon) := -\phi_n + \epsilon \left(\phi_{n+1} + \phi_{n-1} + \gamma {\phi_n}^3\right) = 0, \tag{2.1}$$

with $\phi_0 = A$.

When $|\epsilon|$ is small enough, apart from the boundary, the leading order solution of ϕ_n would formally satisfy

$$\phi_n \left(-1 + \epsilon\gamma {\phi_n}^2\right) \approx 0, \tag{2.2}$$

from which we obtain that $\phi_n \approx 0$ and $\phi_n \approx \pm 1/\sqrt{\epsilon\gamma}$. It physically means that the arrays are almost uncoupled and indicates that solutions of (2.1) can be expressed in terms of an asymptotic or a perturbation expansion in ϵ. It also says that when we consider finitely long waveguide arrays, i.e., $n = 1, 2, \ldots, N$, (2.1) can have at most 3^N solutions. Yet, only some of them are related to the nonlinear tunneling phenomenon presented in Figure 1.1. We are especially interested in solutions with a magnitude that is monotonically decaying with the property $|\phi_n| \to 0$ as $n \to \infty$. This consideration is based on the fact that when the driving frequency is in the forbidden band and the driving amplitude is below the critical threshold, the solution profile is monotonically decaying as $n \to \infty$ (see the top right panel of Figure 1.1). Moreover, we only need to consider particularly a family of *rapidly decaying discrete solitons*, which is defined as follows.

Definition 2.1. *Let $\phi_n = \sum_{k=0}^{\infty} a_{n,k}\vartheta_k(\epsilon)$ be a solution of* (2.1)*, where $n \in \mathbb{Z}^+$, $\vartheta_k(\epsilon)$ is an asymptotic sequence, and $\vartheta_k(\epsilon) = o\left(\vartheta_{k-1}(\epsilon)\right)$ for $\epsilon \to 0$. ϕ_n is a rapidly decaying discrete soliton of* (2.1) *if $|\phi_n|$ is a monotonically decreasing function to* 0 *as $n \to \infty$ with a property that to the leading order $\mathcal{O}(\vartheta_0)$ only the first lattice site is nonzero, i.e., $a_{1,0} \neq 0$ and $a_{n,0} = 0$, $n \neq 1$.*

As an example of this definition, let us consider the following solution of (2.1):

$$\Phi_0(n, A) = \begin{cases} -\dfrac{1}{\sqrt{\gamma}}\left(\dfrac{1}{\sqrt{\epsilon}} + \dfrac{\sqrt{\epsilon}}{2}\right) - A\dfrac{\epsilon}{2} + \mathcal{O}(\epsilon^{3/2}),\ n = 1, \\ \dfrac{1}{\sqrt{\gamma}}\left(\dfrac{1}{\sqrt{\epsilon}} + \dfrac{\sqrt{\epsilon}}{2}\right) + \mathcal{O}(\epsilon^{3/2}),\ n = 2, \\ \mathcal{O}(\sqrt{\epsilon}) \quad \text{otherwise.} \end{cases} \tag{2.3}$$

This solution is obtained from the expansion: $\phi_1 = -1/\sqrt{\epsilon\gamma} + a_{1,1}\sqrt{\epsilon} + a_{1,2}\epsilon + \cdots$, $\phi_2 = 1/\sqrt{\epsilon\gamma} + a_{2,1}\sqrt{\epsilon} + a_{2,2}\epsilon + \cdots$, $\phi_3 = 0 + a_{3,1}\sqrt{\epsilon}+$, and $\phi_n = 0 + \cdots$ for $n > 3$. Substituting the ansatz to (2.1) will yield polynomials in ϵ. Equating the coefficients of the polynomials for all orders of ϵ to zero will yield equations for $a_{k,l}$ that have to be solved simultaneously to obtain (2.3).

It is clear that the profile of $|\Phi_0(n, A)|$ (2.3) is monotonically decaying in n. However, this solution is not rapidly decaying as to the leading order, i.e., $\mathcal{O}(1/\sqrt{\epsilon})$, $|\Phi_0(2, A)| = |\Phi_0(1, A)| \neq 0$.

The existence of rapidly decaying solutions of (2.1) when $A = \mathcal{O}(1)$ is guaranteed by the following theorem.

Theorem 2.2. *Let A be of $\mathcal{O}(1)$. Then for ϵ positive and small there are three rapidly decaying discrete solitons of the static equation* (2.1)*. Denoted by Φ_j, $j = 1, 2, 3$, the solitons are given by*

$$\Phi_1(n, A) = \begin{cases} \dfrac{1}{\sqrt{\epsilon\gamma}} - A\dfrac{\epsilon}{2} - \dfrac{\epsilon^{3/2}}{2\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 1, \\ \dfrac{\sqrt{\epsilon}}{\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 2, \\ \dfrac{\epsilon^{3/2}}{\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 3, \\ 0 + \mathcal{O}(\epsilon^2) \quad \textit{otherwise,} \end{cases} \tag{2.4}$$

$$\Phi_2(n, A) = \begin{cases} A\epsilon + A\epsilon^3 + \mathcal{O}(\epsilon^5),\ n = 1, \\ A\epsilon^2 + \mathcal{O}(\epsilon^4),\ n = 2, \\ A\epsilon^3 + \mathcal{O}(\epsilon^5),\ n = 3, \\ 0 + \mathcal{O}(\epsilon^4) \quad \textit{otherwise,} \end{cases} \tag{2.5}$$

$$\Phi_3(n, A) = \begin{cases} -\dfrac{1}{\sqrt{\epsilon\gamma}} - A\dfrac{\epsilon}{2} + \dfrac{\epsilon^{3/2}}{2\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 1, \\ -\dfrac{\sqrt{\epsilon}}{\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 2, \\ -\dfrac{\epsilon^{3/2}}{\sqrt{\gamma}} + \mathcal{O}(\epsilon^2),\ n = 3, \\ 0 + \mathcal{O}(\epsilon^2) \quad \textit{otherwise.} \end{cases} \tag{2.6}$$

Proof. Because we are looking for rapidly decaying solitons, to the leading order (2.1) can be represented by

$$-\phi_1 + \epsilon A + \gamma\epsilon{\phi_1}^3 = 0. \tag{2.7}$$

Equation (2.7) is a cubic equation similar to (2.2), also with three roots. However, as $\epsilon \to 0$, (2.7) reduces to a linear equation $\phi_1 = 0$ with only a single root. Therefore, finding the roots of the equation is a singular perturbation problem. Following, e.g., [13] (see Example 3 of sections 2.1 and 2.2), one will obtain the roots of (2.7), i.e., $\phi_1 = A\epsilon + \cdots$ and $\phi_1 = \pm 1/\sqrt{\gamma\epsilon} + \cdots$. This concludes that there are three rapidly decaying solutions of (2.1). In the following, let us name the solitons $\Phi_j(n, A)$, $j = 1, 2, 3$, with $\Phi_1(1, A) = 1/\sqrt{\epsilon\gamma} + \cdots$, $\Phi_2(1, A) = \epsilon A + \cdots$, and $\Phi_3(1, A) = -1/\sqrt{\epsilon\gamma} + \cdots$. The existence of $\Phi_j(n, A)$ for (2.1) follows immediately from the implicit function theorem (see, e.g., [14]) since F is differentiable and the Jacobian matrix of problem (2.1) $DF(\phi, 0)$ is invertible. Explicit calculations to obtain (2.4)–(2.6) can be done similarly following the derivation of (2.3). □

If one compares the above theorem and the top right panel of Figure 1.1, it can be recognized immediately that the solution observed in the panel in the limit $z \to \infty$ is nothing else but $|\Phi_2(n, A)|$.

One still can obtain the existence of the above rapidly decaying solutions even when $A \gg 1$, as stated in the following theorem.

THEOREM 2.3. *Let A be scaled to $A = \tilde{A}/\epsilon^{3/2}$, $\tilde{A} < 2/\sqrt{27\gamma}$.*

$$\Phi_j(n, A) = \begin{cases} \dfrac{\Phi_j^0}{\sqrt{\epsilon}} + \dfrac{\tilde{A}}{3\gamma{\Phi_j^0}^2 - 1}\left(\sqrt{\epsilon} - \epsilon\right) + \mathcal{O}(\epsilon^{3/2}), \ n = 1, \\ \Phi_j^0\sqrt{\epsilon} + \mathcal{O}(\epsilon^{3/2}), \ n = 2, \\ 0 + \mathcal{O}(\epsilon^{3/2}) \quad \textit{otherwise,} \end{cases} \tag{2.8}$$

with Φ_j^0 given by

$$\Phi_j^0 = \begin{cases} \dfrac{2}{\sqrt{3\gamma}}\cos\left(\dfrac{1}{3}\arccos\left(\dfrac{-\tilde{A}\sqrt{27\gamma}}{2}\right)\right), & j = 1, \\ \dfrac{2}{\sqrt{3\gamma}}\cos\left(\dfrac{4\pi}{3} + \dfrac{1}{3}\arccos\left(\dfrac{-\tilde{A}\sqrt{27\gamma}}{2}\right)\right), & j = 2, \\ \dfrac{2}{\sqrt{3\gamma}}\cos\left(\dfrac{2\pi}{3} + \dfrac{1}{3}\arccos\left(\dfrac{-\tilde{A}\sqrt{27\gamma}}{2}\right)\right), & j = 3. \end{cases} \tag{2.9}$$

Moreover, if we write $A = 2/\sqrt{(27\gamma\epsilon^3)} - \widehat{A}\sqrt{\epsilon}$, with $\widehat{A} > 1/\sqrt{3\gamma}$, then Φ_j, $j = 1, 2$, can be written as

$$\Phi_{1,2} = \begin{cases} \dfrac{1}{\sqrt{3\gamma}\sqrt{\epsilon}} \mp \sqrt{\dfrac{\widehat{A}}{\sqrt{3\gamma}} - \dfrac{1}{3\gamma}}\sqrt{\epsilon} + \mathcal{O}(\epsilon^{3/2}), \ n = 1, \\ \dfrac{\sqrt{\epsilon}}{\sqrt{3\gamma}} + \mathcal{O}(\epsilon^{3/2}), \ n = 2, \\ \mathcal{O}(\epsilon^{3/2}) \quad \textit{otherwise.} \end{cases} \tag{2.10}$$

Proof. As we are interested in the case of $A \gg 1$, we first scale $A = \tilde{A}/\epsilon^{3/2}$ and correspondingly write $\Phi_j(n, A) = \Phi_j^0(n, A)/\sqrt{\epsilon} + \cdots$, $j = 1, 2, 3$, with $\Phi_j^0(n, A) = 0$

for all $1 < n \in \mathbb{Z}^+$. Substituting the expansion to (2.1) and identifying coefficients for power series of $\mathcal{O}(1/\sqrt{\epsilon})$ yields the following cubic equation for $\Phi_j^0(1, A) = \Phi_j^0$:

$$G\left(\Phi_j^0\right) := -\Phi_j^0 + \tilde{A} + \gamma\left(\Phi_j^0\right)^3 = 0. \tag{2.11}$$

Equation (2.11) cannot be solved perturbatively to obtain the roots Φ_j^0 as before since all the terms in (2.11) are of the same order. Therefore, we need the following lemma on cubic equations.

LEMMA 2.4. *Consider the following polynomial equation:*

$$g(x) = ax^3 + bx^2 + cx + d, \quad a, b, c, d \in \mathbb{R}.$$

Let

$$X = \frac{-b}{3a}, \quad Y = g(X), \quad h = 2av^3,$$
$$v^2 = \frac{b^2 - 3ac}{9a^2}, \quad \theta = \frac{1}{3}\arccos\left(\frac{-Y}{h}\right).$$

If $Y^2 < h^2$, then the cubic equation has three distinct real roots given by

$$x_1 = X + 2v\cos\theta, \tag{2.12}$$
$$x_2 = X + 2v\cos(4\pi/3 + \theta), \tag{2.13}$$
$$x_3 = X + 2v\cos(2\pi/3 + \theta), \tag{2.14}$$

where

$$x_1 > x_2 > x_3.$$

When $Y^2 = h^2$, two of the roots which are neighbors to each other, i.e., x_1 and x_2 or x_2 and x_3, will collide in a saddle-node bifurcation and disappear when $Y^2 > h^2$, i.e., the cubic equation then has only a single real root.

Proof. There is an enormous number of textbooks and online references on cubic equations; see, e.g., http://mathworld.wolfram.com/CubicFormula.html. The expression of the cubic polynomial roots (2.12)–(2.14) is using Nickalls's geometric representation [15]. □

According to Lemma 2.4, (2.11) has geometric representation parameters:

$$X = 0, \quad Y = \tilde{A}, \quad v = \frac{1}{\sqrt{3\gamma}}, \quad h = \frac{2}{\sqrt{27\gamma}}, \quad \theta = \frac{1}{3}\arccos\left(\frac{-Y}{h}\right),$$

from which we can conclude that (2.11) has three real roots when $\tilde{A} < 2/\sqrt{27\gamma}$. The roots of (2.11), i.e., (2.9), are obtained using (2.12)–(2.13). Then the continuation of Φ_j^0 can be obtained immediately using the implicit function theorem.

It is then straightforward to calculate that when $\tilde{A} = 2/\sqrt{27\gamma}$, $\Phi_{1,2}^0 = 1/\sqrt{3\gamma}$ as $2\cos\theta = 2\cos(4\pi/3 + \theta) = 1$. Hence, we know that $\Phi_1(n, A)$ collides with $\Phi_2(n, A)$ in a saddle-node bifurcation.

For the value of A close to the occurrence of the saddle-node bifurcation, we write $A = 2/\sqrt{27\gamma\epsilon^3} - \widehat{A}\sqrt{\epsilon}$. In this case, the implicit function theorem cannot be immediately employed to prove the existence of Φ_1 and Φ_2 as we need a bound for $\widehat{A}$.

First, we substitute $\Phi_j = \Phi_j^0(n,A)/\sqrt{\epsilon} + \sqrt{\epsilon}\Phi_j^1(n,A)$, $j = 1,2$, in the steady state equation (2.1) with $\Phi_j^0(n,A) = 1/\sqrt{3\gamma}$ for $n = 1$ and 0 otherwise. This then gives the following equations:

$$\begin{aligned}
&\tilde{G}_1(\Phi_j^1,\epsilon) := \Phi_j^1(2,A) - \widehat{A} + \sqrt{3\gamma}\left(\Phi_j^1(1,A)\right)^2 + \epsilon\gamma\left(\Phi_j^1(1,A)\right)^3 = 0,\\
&\tilde{G}_2(\Phi_j^1,\epsilon) := -\Phi_j^1(2,A) + \epsilon\Phi_j^1(3,A) + \frac{1}{\sqrt{3\gamma}} + \epsilon\Phi_j^1(1,A) + \epsilon^2\gamma\left(\Phi_j^1(1,A)\right)^3 = 0,\\
&\tilde{G}_n(\Phi_j^1,\epsilon) := \Phi_j^1(n,A) + \epsilon\left(\Phi_j^1(n+1,A) + \Phi_j^1(n-1,A) + \epsilon\gamma\Phi_j^1(n,A)^3\right) = 0, \quad n \neq 1,2.
\end{aligned}$$

Taking $\epsilon = 0$, the above equations give us

$$\begin{aligned}
\Phi_j^1(1,A) &= \pm\sqrt{\frac{\widehat{A}}{\sqrt{3\gamma}} - \frac{1}{3\gamma}},\\
\Phi_j^1(2,A) &= \frac{1}{\sqrt{3\gamma}},\\
\Phi_j^1(n,A) &= 0, \quad n \neq 1,2.
\end{aligned}$$

Note that the $\pm$-solutions collide for $\widehat{A} = 1/\sqrt{3\gamma}$. Because the linearization $D\tilde{G}(\Phi_j^1,0)$ is invertible for $\widehat{A} > 1/\sqrt{3\gamma}$, the implicit function theorem can be applied again and we have the existence of rapidly decaying solitons $\Phi_j = \Phi_j^0(n,A)/\sqrt{\epsilon} + \sqrt{\epsilon}\Phi_j^1(n,A)$, $j = 1,2$. □

With this theorem, we then have shown that Φ_1 collides in a saddle-node bifurcation with Φ_2. Yet, we cannot directly claim that this is the source of the supratransmission observed in Figure 1.1 before we show and discuss the stability of the two solitons.

2.2. Stability analysis. After discussing the existence of rapidly decaying solitons of (2.1), we study their stability. If ϕ_n, $n = 1,2,\ldots$, is a solution of (2.1), then the linear spectral stability of ϕ_n can be obtained by substituting the ansatz $\psi_n = (\phi_n + \delta[v_n e^{i\lambda z} + \overline{w_n}e^{-i\overline{\lambda}z}])e^{i\Delta z}$ with $\lambda \in \mathbb{C}$, $(v_n,w_n) \in \mathbb{C}^2$, and $n \in \mathbb{Z}^+$ into (1.1). Linearizing the equation to $\mathcal{O}(\delta)$, we obtain the eigenvalue problem

$$\lambda\epsilon\begin{pmatrix} v_n \\ w_n \end{pmatrix} = \epsilon\sigma\begin{pmatrix} v_{n-1} \\ w_{n-1} \end{pmatrix} + \mathcal{L}\begin{pmatrix} v_n \\ w_n \end{pmatrix} + \epsilon\sigma\begin{pmatrix} v_{n+1} \\ w_{n+1} \end{pmatrix}, \tag{2.15}$$

with

$$\begin{aligned}
\begin{pmatrix} v_0 \\ w_0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \sigma = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},\\
\mathcal{L} &= \begin{pmatrix} -1 + 2\epsilon\gamma|\phi_n|^2 & \epsilon\gamma{\phi_n}^2 \\ -\epsilon\gamma{\phi_n}^2 & 1 - 2\epsilon\gamma|\phi_n|^2 \end{pmatrix}, \quad n \in \mathbb{Z}^+,
\end{aligned}$$

where we have scaled $\Delta \to 1$.

The natural domain for $\tilde{\mathcal{L}} = (\epsilon\sigma \quad \mathcal{L} \quad \epsilon\sigma)$ is $L^2(\mathbb{C})$. We call λ an eigenvalue of $\tilde{\mathcal{L}}$ if there is a function $\{v_n\}_{n\in\mathbb{Z}^+}, \{w_n\}_{n\in\mathbb{Z}^+} \in L^2(\mathbb{C})$ which satisfies (2.15). Since $\tilde{\mathcal{L}}$ depends smoothly on A, the eigenvalues of $\tilde{\mathcal{L}}$ will depend smoothly on A, too. ϕ_n is linearly stable if the imaginary part of λ is zero, i.e., $\mathrm{Im}(\lambda) = 0$.

The continuous spectrum is obtained by substituting

$$v_n = Ae^{ikn}, \quad w_n = Be^{ikn}, \quad \phi_n = 0$$

in (2.15), from which we will obtain

$$\epsilon\lambda = \pm 2\epsilon \cos k \mp 1.$$

Thus, the continuous spectrum of solutions under investigation is the range

$$\lambda \in \left(-\frac{1}{\epsilon} - 2, -\frac{1}{\epsilon} + 2\right) \quad \text{and} \quad \lambda \in \left(\frac{1}{\epsilon} - 2, \frac{1}{\epsilon} + 2\right). \tag{2.16}$$

As the continuous spectrum lies in the real axis, the stability of the solutions is determined only by the discrete spectrum, i.e., eigenvalues. For the solutions given in Theorems 2.2 and 2.3, we have the following stability results.

THEOREM 2.5. *For small driving amplitude $A = \mathcal{O}(1)$, the various rapidly decaying discrete solitons have the following properties:*

1. *The discrete soliton Φ_1 is unstable. It has a single imaginary eigenvalue.*
2. *The soliton Φ_2 is strictly stable as the soliton has no discrete eigenvalues.*
3. *The discrete soliton Φ_3 is stable. It has a single real eigenvalue.*

Proof. We are looking for eigenvectors that are also rapidly decaying. Therefore, the eigenvalue problem (2.15) to the leading order can be approximated by the linear eigenvalue problem

$$\lambda\epsilon \begin{pmatrix} v_1 \\ w_1 \end{pmatrix} = \mathcal{L} \begin{pmatrix} v_1 \\ w_1 \end{pmatrix},$$

which gives the following approximate eigenvalues:

$$\lambda = \pm\frac{1}{\epsilon}\sqrt{3\left(\epsilon\gamma{\phi_n}^2\right)^2 - 4\epsilon\gamma{\phi_n}^2 + 1}. \tag{2.17}$$

In the above expression, we have taken into account the fact that $\phi_n \in \mathbb{R}$.

For the stability of Φ_1 and Φ_3, substitute $\phi_1 = \Phi_j(1, A)$, $j = 1, 3$, into (2.17). Taking the series expansion of the expression gives the following eigenvalue λ for $\Phi_j(n, A)$:

$$\lambda = \begin{cases} \epsilon^{-1/4}\sqrt{2A\sqrt{\gamma}}i + \mathcal{O}(\epsilon^{1/4}), & j = 1, \\ \epsilon^{-1/4}\sqrt{2A\sqrt{\gamma}} + \mathcal{O}(\epsilon^{1/4}), & j = 3. \end{cases} \tag{2.18}$$

Because the eigenvalue of $\Phi_1(n, A)$ has a nonzero imaginary part, we conclude that to the leading order Φ_1 is unstable, as opposed to Φ_3.

As for $\phi_1 = \Phi_2(1, A)$, the series expansion of (2.17) gives

$$\lambda = 1/\epsilon + \mathcal{O}(\epsilon^2). \tag{2.19}$$

Because λ is inside the continuous spectrum (2.16), our assumption that the eigenfunction is rapidly decaying is not justified. Nonetheless, we know that Φ_2 bifurcates from a uniform solution $\phi_n \equiv 0$, which is stable. Because L depends smoothly on A, we then can conclude that Φ_1 has no eigenvalue. □

When the driving amplitude is large, we also have the following theorem.

THEOREM 2.6. *For large driving amplitude $A = \tilde{A}\epsilon^{-3/2}$, the various rapidly decaying discrete solitons have the following properties:*

1. *The discrete soliton Φ_1 is unstable with a single imaginary eigenvalue.*
2. *The soliton Φ_2 is strictly stable with a single real eigenvalue.*

3. *The discrete soliton Φ_3 is in general stable with a single eigenvalue, except in a finite interval, where our asymptotic analysis is inconclusive.*

To the leading order, the eigenvalue of the three solitons is given by

$$\lambda = K/\epsilon + \frac{\tilde{A}}{\Phi_j^0\left(3\gamma\left(\Phi_j^0\right)^2 - 1\right)}\left(\frac{\left(3\gamma^2\left(\Phi_j^0\right)^4 - 1\right)}{K} + K\right) + \mathcal{O}(\sqrt{\epsilon}), \tag{2.20}$$

with $K = \sqrt{3(\gamma{\Phi_j^0}^2)^2 - 4\gamma{\Phi_j^0}^2 + 1}$. Moreover, by writing $A = 2/\sqrt{27\gamma e^3} - \widehat{A}\sqrt{\epsilon}$, the eigenvalue of $\Phi_{1,2}$ is given by

$$\lambda = \frac{2}{3^{1/4}\sqrt{\epsilon}}\sqrt{\mp\sqrt{\widehat{A}\sqrt{\frac{\gamma}{3}} - \frac{1}{3}}} + \mathcal{O}(\sqrt{\epsilon}), \tag{2.21}$$

with the minus sign for the eigenvalue of Φ_1 and the plus sign for Φ_2.

Proof. The proof of Theorem 2.6 is similar to the proof of Theorem 2.5. The stability result of Φ_3 cannot be deduced immediately because the expression of Φ_3 is not trivial. The presence of a finite interval where our asymptotic analysis is inconclusive cannot be seen clearly. It is inconclusive because there is a range of A in which λ is in the domain of the continuous spectrum (2.16). A numerical proof will be presented in the following section. □

2.3. Analysis for the case of $\Delta < -2$. We omit the details and the rigorous proof, but it can be shown that for $\Delta < -2$, there is only one rapidly decaying soliton which is stable for any driving amplitude. The idea is as follows.

Instead of (2.1), consider

$$\phi_n = -\epsilon(\phi_{n+1} + \phi_{n-1}) - \gamma\epsilon{\phi_n}^3, \quad \phi_0 = A, \tag{2.22}$$

where we again have scaled $0 > \Delta \to 1$ and define $\epsilon = 1/|\Delta|$. For a rapidly decaying solution, the leading order equation of (2.22) is then given by

$$f := \phi_1 + \epsilon\left(A + \gamma{\phi_1}^3\right) = 0. \tag{2.23}$$

It is clear that $f \to \pm\infty$ as $\phi \to \pm\infty$. Yet, f has no critical point, i.e., $df/d\phi_1 > 0$. Therefore, one can conclude that f is a monotonically increasing function which intersects the horizontal axis once, i.e., f has one real root. The stability of this rapidly decaying solution might be determined immediately following Theorems 2.5 and 2.6. Our numerics, which are not presented here, show that when $A = \mathcal{O}(1)$ the solution is stable with no discrete spectrum, and when $A = \mathcal{O}(1/\epsilon^{3/2})$ there is an eigenvalue bifurcating from the upper edge of the continuous spectrum. Hence, the soliton is stable all the way to $A \to \infty$, which explains why there is no supratransmission for $\Delta < -2$.

2.4. Numerical results. To accompany our analytical results, we have used numerical calculations. For that purpose, we have made a continuation program based on a Newton iteration technique to obtain stationary rapidly decaying discrete solitons of (2.1) and an eigenvalue problem solver to solve (2.15) in MATLAB. Throughout the subsection, we consider in particular $\Delta = 10$. Even though there is no prominent supratransmission of energy for this value of Δ, it is taken solely as an example to show that especially in the regime of Δ large, our asymptotic analysis explains the

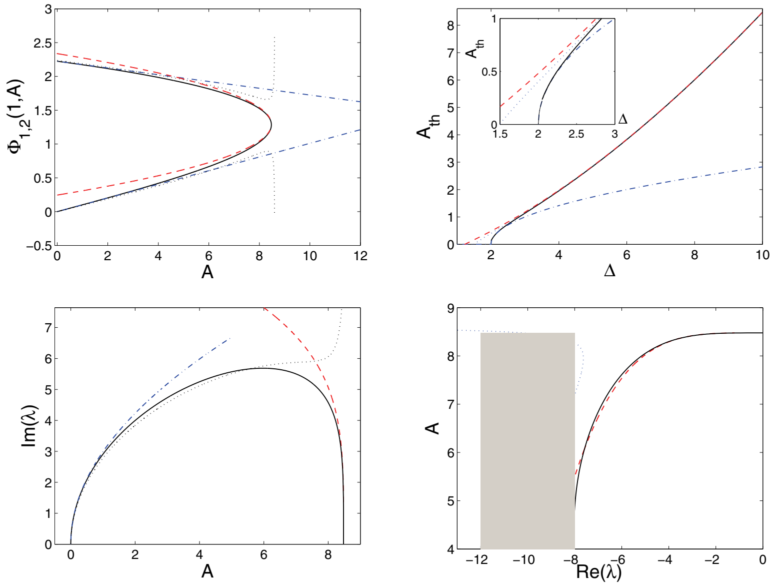


FIG. 2.1. *Presented is the comparison between the numerically obtained results and the analytical calculations presented in section* 2. *Top left panel: the existence curve of* Φ_1 *and* Φ_2 *represented by the solution of the first site, where the upper and lower branches correspond to the existence curve of* Φ_1 *and* Φ_2*, respectively. Top right panel: the threshold amplitude* A_{th} *as a function of the propagation constant* Δ*. Bottom panels: the critical eigenvalue of* Φ_1 *(left) and* Φ_2 *(right) as a function of the driving amplitude* A*. Shaded region in the bottom right panel shows the region for the continuous spectrum. Analytical approximations calculated in section* 2 *are also presented as dashed, dotted, and dash-dotted lines (see the text).*

problem well. It will be shown below that, e.g., even using the first two terms of the approximate threshold amplitude, our analytical result is already relatively in agreement with the numerical results.

We summarize our results and discussions for the existence and the stability of Φ_1 and Φ_2 in Figure 2.1. At the top left panel of the figure, we present the existence of Φ_j, $j = 1, 2$, represented by the solution of the first site, where the upper and lower branches correspond to the existence curve of Φ_1 and Φ_2, respectively. Numerical results are represented by the solid lines. Our analytical result as given by (2.4) and (2.5) which is supposed to be valid when $A = \mathcal{O}(1)$ is depicted by dash-dotted lines. As for the analytical approximations for $A = \mathcal{O}(1/\epsilon^{3/2})$, i.e., (2.8) and (2.10), they are presented as dotted and dashed lines, respectively. It is interesting to note that Figure 2.1 shows clearly a good agreement between our analytical and the numerical results.

Top right panel of Figure 2.1 presents the comparison between the critical amplitude $A_{\mathrm{th}}(\Delta)$ calculated numerically from (2.1) and our approximation $A_{\mathrm{th}}(\Delta) = 2/\sqrt{27\gamma\epsilon^{3/2}} - \sqrt{\epsilon}/\sqrt{3\gamma}$ (see Theorem 2.3), which are presented in solid and dashed lines, respectively. The numerical results were also checked against the full dynamics of the original problem (1.1), where an agreement is obtained as it should be provided that τ is large enough. Note the good agreement when $\Delta \gg 1$. As a comparison with the analytical approximation obtained by Khomeriki [5], we also present $A_{\mathrm{th}}(\Delta) = \sqrt{\Delta - 2}$ as a dash-dotted line.

It is interesting to note that in the limit $\Delta \to 2$ our analytical approximation does not diverge. As is shown in the inset of the top right panel, the difference of the approximate value of the threshold amplitude and the numerical result at $\Delta = 2$ is about 50%. Using the same method presented in the preceding sections, we obtain that the first three terms of the approximation of $A_{\text{th}}(\Delta)$ are actually given by

$$A_{\text{th}}(\Delta) = \frac{2}{\sqrt{27\gamma\epsilon^{3/2}}} - \frac{\sqrt{\epsilon}}{\sqrt{3\gamma}} - \frac{13\sqrt{3}}{36\sqrt{\gamma}}\epsilon^{5/2}. \tag{2.24}$$

The plot of this curve is depicted in the same panel as a dotted line, where one can see that the difference now has decreased by about 10%. This then motivates us to question whether the infinite power series of the approximate threshold amplitude $A_{\text{th}}(\Delta)$ is actually convergent uniformly to the critical amplitude curve. Considering the fact that the region of interest is on $0 < \epsilon \leq 1/2$ and the coefficients of the power series are so far bounded, the answer might well be affirmative. Yet, this question is out of the scope of the present paper and will therefore be addressed in future investigations.

After presenting the numerical and the analytical results for the existence of Φ_1 and Φ_3, next we consider the stability of the solitons. The bottom panels of Figure 2.1 present the comparison between the results. The left panel shows the imaginary part of the critical eigenvalue of Φ_1 as a function of A in its existence region. It is clear that the soliton is always unstable. The right panel presents the eigenvalue of Φ_2 as a function of the driving amplitude, where one can see that the soliton is always stable, as opposed to Φ_1. Our analytical approximations (2.18), (2.20), and (2.21) are presented as well in the two panels as dash-dotted, dotted, and dashed lines, respectively. It is also interesting to note that, as is predicted by Theorem 2.2, Φ_2 has no eigenvalue when A is small. Our analytical approximation (2.21) predicts very well the appearance of the eigenvalue of Φ_2.

Because it is known that Φ_1 is unstable in its entire existence region, it is of interest to see the dynamics with regards to instability. In Figure 2.2 we present the evolution of Φ_1 for a parameter value $A \equiv 8.46$ (A is already at this value from the beginning $z = 0$, as opposed to Figure 1.1, where A is 0 in the beginning and gradually increases to a constant). The top left panel presents the dynamics of Φ_1 with the initial condition $\psi_n(z = 0) = \Phi_1(n, A) - 10^{-4}$. The initial condition $\Phi_1(n, A)$ is obtained numerically from (2.1). The top right panel depicts the behavior of the first site in time, where one can see that the instability manifests in the form of the soliton's oscillations. Interestingly, if we start with an initial condition of the form $\psi_n(z = 0) = \Phi_1(n, A) + 10^{-4}$, the solution has a similar instability behavior but with a different oscillation maximum. The dynamics are presented in the bottom panels of Figure 2.2. It is important to note that with such a small change, the dynamics can be significantly different. This duality therefore might be employed as a small intensity light detector similar to the proposal of [12].

We have numerically analyzed as well the existence and the stability of the soliton Φ_3. We summarize our results in Figure 2.3. The numerical result for the existence of the soliton is shown in the top left panel of the figure. Our analytical approximations (2.6) and (2.8) are shown in dash-dotted and dotted lines, respectively, where one can see the good agreement between the numerical and the analytical results.

After studying the existence of the discrete soliton, we next present our stability analysis of the soliton. Shown in the bottom panels of Figure 2.3 is the numerically obtained critical eigenvalue of Φ_3 as a function of A. In the bottom left panel is the real part of the critical eigenvalue. It is clear that when $A = 0$, the eigenvalue is a double

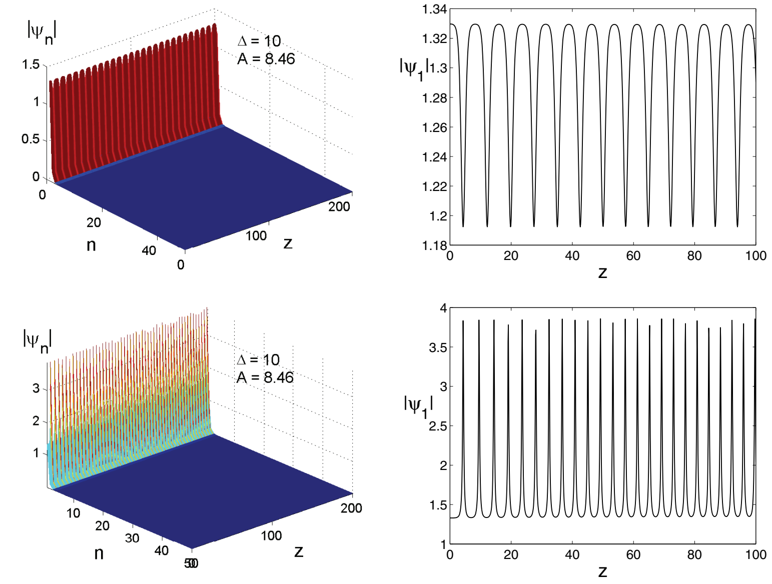


FIG. 2.2. *The instability dynamics of Φ_1 for $A = 8.46$ and $\Delta = 10$. It is presented that even with a tiny change in perturbation the dynamics can be significantly different (see the text).*

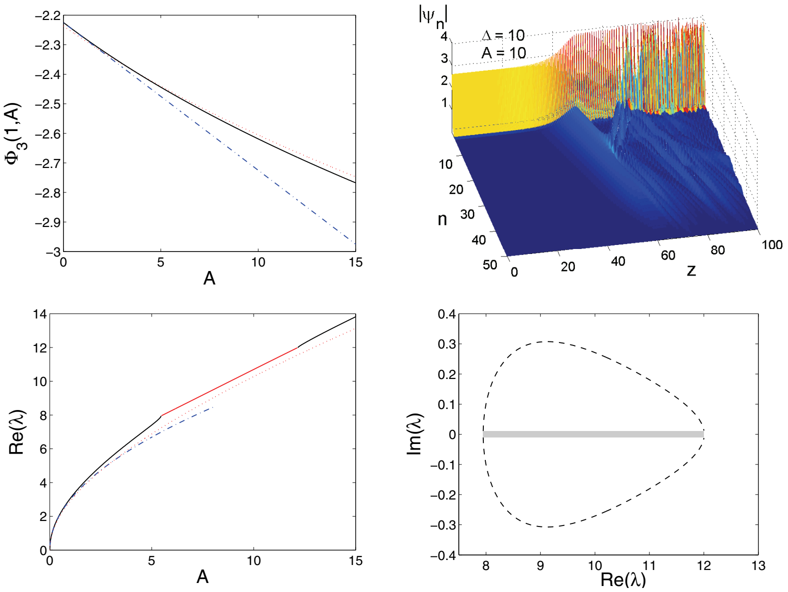


FIG. 2.3. *Similar to Figure 2.1 but for the soliton Φ_3. Top left panel shows the numerical results for the existence of Φ_3 versus the driving amplitude A (solid line). Presented is the value of the solution at the first site, i.e., $\Phi_3(1, A)$. The bottom left panel presents the stability of the soliton. The red solid line that separates the black solid line indicates that the soliton is unstable in this region. The behavior of the critical eigenvalue in the complex plane is depicted in the bottom right panel. In the panel, the parametric variable is the driving parameter A. The top right panel shows the dynamics of the soliton when it is unstable. See the text for the analytical approximation curves. (For interpretation of the references to color in this figure legend, the reader is referred to the online version of this article.)*

eigenvalue at zero. As soon as A is increased, the zero eigenvalue bifurcates along the real line. At a critical driving amplitude, the eigenvalue collides with the lower boundary of the continuous spectrum. The result of the collision is the bifurcation of the eigenvalue into the complex plane resulting in an eigenvalue with nonzero imaginary part. In the bottom right panel, we present the trajectory of the eigenvalue in the complex plane as A is increased. One can then see that there is also another critical amplitude above which the eigenvalue becomes real again, i.e., the soliton becomes stable. In the region where the imaginary part is nonzero, we depict the curve in the bottom left panel of Figure 2.3 in solid red line. We also compare it with our analytical approximations, (2.18) and (2.20), which are shown as dash-dotted and dotted lines, respectively. In Theorem 2.6, it is stated that our analytical approximation is inconclusive for the case of A large. As can be seen from Figure 2.3, our analytical approximation equation (2.20) is always real. This is because when the real part of the eigenvalue is in the region of the continuous spectrum, our assumption that the eigenfunction is rapidly decreasing is not justified.

It is then interesting to see the dynamics of the instability. In the top right panel, we depict the evolution of an unstable discrete soliton of type Φ_3. The parameter values are depicted in the figure. The setup for the driving amplitude is similar to the setup of Figure 2.2.

Regarding the involvement of Φ_3 in the dynamics of the driven boundary waveguides (1.1) (see Figure 1.1), it is not clear, when Φ_2 disappears, whether it evolves into Φ_3.

3. Conclusions. We have analyzed mathematically the mechanism of supratransmissions observed in a boundary driven discrete nonlinear Schrödinger equation describing electromagnetic fields in waveguide arrays. We have shown that the source of the phenomenon is the presence of a saddle-node bifurcation between a stable discrete soliton and an unstable one. We have shown as well numerically that the unstable one can exhibit a different dynamics, sensitive to the perturbation. We therefore argue that it might be possible to propose it as a weak signal light detector.

Acknowledgments. The author wishes to thank Panayotis Kevrekidis and Ramaz Khomeriki for numerous useful interactions and discussions. The manuscript has greatly benefited from the constructive comments and suggestions of two anonymous reviewers.

REFERENCES

[1] F. Geniet and J. Leon, *Energy transmission in the forbidden band gap of a nonlinear chain*, Phys. Rev. Lett., 89 (2002), article 134102.

[2] S. Savel'ev, A. L. Rakhmanov, V. A. Yampol'skii, and F. Nori, *Analogues of nonlinear optics using terahertz Josephson plasma waves in layered superconductors*, Nat. Phys., 2 (2006), pp. 521–525.

[3] J. E. Macías-Díaz and A. Puri, *On the propagation of binary signals in damped mechanical systems of oscillators*, Phys. D, 228 (2007), pp. 112–121.

[4] S. Savel'ev, V. A. Yampol'skii, A. L. Rakhmanov, and F. Nori, *Layered superconductors as nonlinear waveguides for terahertz waves*, Phys. Rev. B, 75 (2007), article 184503.

[5] R. Khomeriki, *Nonlinear band gap transmission in optical waveguide arrays*, Phys. Rev. Lett., 92 (2004), article 063905.

[6] Y. S. Kivshar and M. Peyrard, *Modulational instabilities in discrete lattices*, Phys. Rev. A, 46 (1992), pp. 3198–3205.

[7] T. R. Melvin, A. R. Champneys, P. G. Kevrekidis, and J. Cuevas, *Radiationless traveling waves in saturable nonlinear Schrödinger lattices*, Phys. Rev. Lett., 97 (2006), article 124101.

[8] R. Khomeriki, S. Lepri, and S. Ruffo, *Nonlinear supratransmission and bistability in the Fermi-Pasta-Ulam model*, Phys. Rev. B, 70 (2004), article 066626.

[9] J. Leon, *Nonlinear supratransmission as a fundamental instability*, Phys. Lett. A, 319 (2003), pp. 130–136.

[10] L. Hadžievski, A. Maluckov, M. Stepić, and D. Kip, *Power controlled soliton stability and steering in lattices with saturable nonlinearity*, Phys. Rev. Lett., 93 (2004), article 033901.

[11] H. Susanto and N. Karjanto, *Calculated threshold of supratransmission phenomena in waveguide arrays with saturable nonlinearity*, J. Nonlinear Optim. Phys. Mater., 17 (2008), pp. 159–165.

[12] R. Khomeriki and J. Leon, *Light detectors bistable nonlinear waveguide arrays*, Phys. Rev. Lett., 94 (2005), article 243902.

[13] A. H. Nayfeh, *Introduction to Perturbation Techniques*, Wiley-Interscience, New York, 1981.

[14] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, Courant Institute, New York, 1974.

[15] R. W. D. Nickalls, *A new approach to solving the cubic: Cardan's solution revealed*, The Math. Gazette, 77 (1993), p. 354.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 126–148

THE SHAPE OF CHARGED DROPS OVER A SOLID SURFACE AND SYMMETRY-BREAKING INSTABILITIES[*]

M. A. FONTELOS[†] AND U. KINDELÁN[‡]

Abstract. We study the static shape of charged drops of a conducting fluid placed over a solid substrate, surrounded by a gas, and in absence of gravitational forces. The question can be formulated as a variational problem where a certain energy involving the areas of the solid-liquid interface and of the liquid-gas interface, as well as the electric capacity of the drop, has to be minimized. As a function of two parameters, Young's angle θ_Y and the potential at the drop's surface V_0, we find the axisymmetric minimizers of the energy and describe their shape. We also discuss the existence of symmetry-breaking bifurcations such that, for given values of θ_Y and V_0, configurations for which the axial symmetry is lost are energetically more favorable than axially symmetric configurations. We prove the existence of such bifurcations in the limits of very flat and almost spherical equilibrium shapes. All other cases are studied numerically with a boundary integral method. One conclusion of this study is that axisymmetric drops cannot spread indefinitely by introducing sufficient amount of electric charges, but can reach only a limiting (saturation) size, after which the axial symmetry would be lost and finger-like shapes energetically preferred.

Key words. electrowetting, symmetry-breaking bifurcations, boundary integral method, variational formulation

AMS subject classifications. 35J50, 65N83, 76D45

DOI. 10.1137/080713707

1. Introduction.

The determination of the stationary shapes of liquid drops surrounded by a vapor phase and in contact with a solid surface is an old problem both in fluid mechanics and in the theory of partial differential equations (see [7] and the references therein). The problem can be posed, since Gauss, in a variational setting consisting of obtaining the configurations of a given mass of fluid that minimize (or in general make extremal) an energy defined by

$$E = \gamma_{lv} A_{lv} - (\gamma_{sv} - \gamma_{sl}) A_{sl} + E_F, \tag{1.1}$$

where γ_{lv}, γ_{sv}, and γ_{sl} denote the liquid-vapor, solid-vapor, and solid-liquid surface tensions, respectively; A_{lv} and A_{sl} denote the area of the liquid-vapor and solid-liquid interfaces, respectively (see Figure 1.1). E_F is the contribution of external forces to the total energy. If the drop is affected by gravity, then $E_F = \int_\Omega \mathbf{g} \cdot \mathbf{x} dx$, where $\mathbf{g}$ is the gravitational force and Ω the domain occupied by the fluid. In absence of external forces, the configurations that minimize the energy (1.1) are spherical caps such that the contact angle θ_Y, called Young's angle, between the liquid-vapor and solid-liquid interfaces satisfies

$$\cos\theta_Y = \frac{\gamma_{sv} - \gamma_{sl}}{\gamma_{lv}}.$$

When the volume of fluid under consideration is sufficiently small, the contribution of gravitational forces to the energy is negligible in comparison with interfacial

[*] Received by the editors January 18, 2008; accepted for publication (in revised form) May 30, 2008; published electronically August 27, 2008.
http://www.siam.org/journals/siap/69-1/71370.html

[†] Instituto de Ciencias Matemáticas (ICMAT, CSIC-UAM-UCM-UC3M), C/ Serrano 123, 28006 Madrid, Spain (marco.fontelos@uam.es).

[‡] Departamento de Matemática Aplicada y Métodos Informáticos, Universidad Politécnica de Madrid, C/ Ríos Rosas 21, 28003 Madrid, Spain (ultano.kindelan@upm.es).

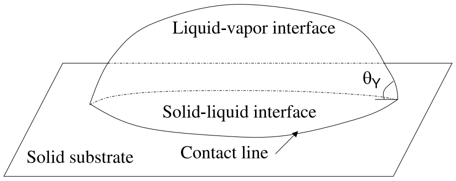


Fig. 1.1. *Sketch of the problem.*

energies. A consistent approximation is then to ignore gravity in (1.1), as it is done systematically in the study of multiphase flows in microfluidic applications, for instance (see [18]). It is precisely in connection with such microfluidic applications that electric fields are incorporated with the purpose of controlling the shape and motion of small masses of fluid. This is the case, for instance, of electrowetting applications in which the shape of a mass of electrically conducting fluid is controlled via the addition of electric charges and application of external electric fields (see [12] and the references therein).

The simplest situation corresponds to a drop of perfectly conducting fluid with a total charge Q. In this case, the energy would be given (cf. [12]) by

$$E = \gamma_{lv}\left[A_{lv} - (\cos\theta_Y)A_{sl}\right] - \frac{1}{2}\varepsilon_0 \int_{\mathbb{R}^3\setminus\Omega} |\mathbf{E}|^2\, dx, \tag{1.2}$$

where $\mathbf{E} = -\nabla V$ is the electric field (V is the electric potential) created by the charges, concentrated in $\partial\Omega$ with a density $\sigma = -\varepsilon_0 \frac{\partial V}{\partial n}$ in a perfect conductor. ε_0 is the dielectric constant of the medium surrounding the drop. The potential V at the surface of a conductor is constant and, in absence of charges in $\mathbb{R}^3 \setminus \Omega$, is harmonic in the exterior of Ω. Hence, V is solution of the boundary value problem

$$\Delta V = 0 \quad \text{at } \mathbb{R}^3 \setminus \Omega, \tag{1.3}$$

$$V = V_0 \quad \text{at } \partial\Omega, \tag{1.4}$$

$$V = O(|\mathbf{r}|^{-1}) \quad \text{as } |\mathbf{r}| \to \infty. \tag{1.5}$$

The electrical energy term in (1.2) can be written, after integration by parts, in the equivalent forms

$$\frac{1}{2}\varepsilon_0 \int_{\mathbb{R}^3\setminus\Omega} |\mathbf{E}|^2\, dx = \frac{1}{2}\varepsilon_0 \int_{\partial\Omega} V\frac{\partial V}{\partial n} dS = \frac{1}{2}\int_{\partial\Omega} V_0\sigma dS = \frac{1}{2}QV_0 = \frac{1}{2}CV_0^2,$$

where C is the capacity of Ω defined as

$$C = -\frac{\varepsilon_0}{V_0}\int_{\Omega} \frac{\partial V}{\partial n} dS.$$

The determination of the capacity of a given set is, in general, a difficult problem. There are explicit expressions for only a few configurations, such as spheres and discs (see [10]). The best source concerning estimation of the capacity of arbitrary sets is [14] and the related article [15]. More complex configurations, such as spherical caps, have a capacity that can be estimated from above and below but no explicit formulae. This is the main difficulty in the deduction of minimizers of (1.2).

A particular case of the problem corresponds to $\theta_Y = 2\pi$, which would correspond to absence of contact with solids. This is the case of levitating droplets, in which the energy is, instead of (1.2),

$$E = \gamma_{lv} A_{lv} - \frac{1}{2} C V_0^2,$$

and extremal values are reached by spheres, as Lord Rayleigh showed in 1882 [17]. This particular situation points to the possibility of appearance of instabilities breaking the axial symmetry and leading to singularities at the drop's surface. When the electric charge (and V_0 therefore) is large enough, those spheres cannot be stable. Two scenarios are possible then: evolution towards nonspherical stationary configurations that accommodate such amount of charge, or destabilization leading to singularities at the interface in finite time. The existence of nonspherical equilibrium configurations in the form of spheroids was proved in [8], and they were characterized as branches of solutions bifurcating, via the Crandall–Rabinowitz theorem, from the branch of spherical solutions. The appearance of singularities in the evolution of initially spherical levitating drops with large charge was shown in [3]. The drop evolves initially into a prolate spheroid with its poles increasing in curvature and becoming conical tips in finite time.

By computing the first variation of the functional (1.2) one arrives at the following equation:

$$\gamma\kappa - \frac{\sigma^2}{2\varepsilon_0} = -p, \tag{1.6}$$

where $\gamma = \gamma_{lv}$, κ is the mean curvature (sum of the principal curvatures) of the liquid-vapor interface at a point $\mathbf{x}$, $\sigma = -\varepsilon_0 \frac{\partial V}{\partial n}$ is the surface charge density at the same point $\mathbf{x}$, and p is a constant to be determined through the constraint that the drop has a given volume. In the fluid dynamics context, p is the difference of pressure across the interface. Equation (1.6) has to be complemented with the boundary condition stating that the normal vector to the interface forms a constant angle with the normal to the solid substrate at any point of the contact line Γ, the set where the liquid-vapor and the solid-liquid interfaces meet. This angle has to be, exactly, θ_Y (cf. [13]). Finally, we introduce characteristic length $(\mathit{Vol.})^{\frac{1}{3}}$, where *Vol.* is the volume of the drop, characteristic potential $(\mathit{Vol.})^{\frac{1}{6}}(\gamma\varepsilon_0^{-1})^{\frac{1}{2}}$, characteristic surface charge density $(\mathit{Vol.})^{-\frac{1}{6}}(\gamma\varepsilon_0)^{\frac{1}{2}}$, and characteristic pressure $\gamma(\mathit{Vol.})^{-\frac{1}{3}}$. Accordingly, we change variables and unknowns in the form $\mathbf{x} \to (\mathit{Vol.})^{\frac{1}{3}}\mathbf{x}$, $V \to (\mathit{Vol.})^{\frac{1}{6}}(\gamma\varepsilon_0^{-1})^{\frac{1}{2}}V$, $\sigma \to (\mathit{Vol.})^{-\frac{1}{6}}(\gamma\varepsilon_0)^{\frac{1}{2}}\sigma$, $p \to \gamma(\mathit{Vol.})^{-\frac{1}{3}}p$ so that space coordinates, potential, surface charge density, and pressure are now dimensionless. We end up with the following dimensionless version of (1.6):

$$\kappa - \frac{\sigma^2}{2} = -p, \tag{1.7}$$

and the variational problem associated to the functional

$$E = [A_{lv} - (\cos\theta_Y)A_{sl}] - \frac{1}{2} C V_0^2, \tag{1.8}$$

where $C = -\frac{1}{V_0}\int_\Omega \frac{\partial V}{\partial n} dS$, V being the solution to (1.3)–(1.5), and Ω is now a domain of unit volume.

One important motivation for this work is its relation with the phenomenon of electrowetting. This consists in the control of the wetting properties of fluids by means of electric fields. The simplest situation is that of a drop of conducting fluid connected to a battery and, therefore, kept at a given difference of potential with respect to an electrode placed at some distance below the solid substrate. In the situation studied in the present paper, such an electrode would be placed at infinity, so that we establish a given potential on the drop's surface and assume the potential to decay at infinity. The simplest theories on electrowetting follow the original ideas of Lippmann, who developed a formula (cf. [11]) that predicts an unlimited spreading of droplets through application of sufficiently strong electrostatic potentials. Nevertheless, the physical observation is that drops do not spread infinitely but reach a saturation regime such that an increase of the potential does not produce any additional spreading but rather the appearance of instabilities at the contact line with subsequent emission of a varying number of satellite filaments (see [12] and the references therein). The demonstration of such facts, also appearing when the electrode is at a fixed distance to the drop, requires a somewhat different analysis and will be published elsewhere.

In this paper we shall study, as a function θ_Y and V_0, the equilibrium configurations. We will present explicit formulae for the geometry of axially symmetric profiles in certain limits of θ_Y and V_0. We will deduce a first constraint to indefinite spreading due to the fact that static solutions develop dewetted cores with the fluid concentrated in a rim around these cores. The second constraint concerns stability. By analyzing the energy functional (1.8) we will conclude that axially symmetric solutions must become unstable under nonaxisymmetric perturbations with n undulations ($n = 2, 3, \ldots$), provided V_0 is large enough. We shall determine numerically (and analytically in some limiting cases) for what values of V_0 such instabilities do develop as a function of θ_Y. This offers a possible explanation to the saturation effect explained above and the contact line instabilities observed in experiments.

The paper is organized as follows. In section 2 we study the radially symmetric configurations. We perform an analysis in the limiting cases of almost spherical drops and almost flat drops and then develop a numerical code to compute the profiles in all intermediate cases. This allows us to represent in a phase diagram the radius of spreading of a drop as a function of V_0 for arbitrary θ_Y. In section 3 we study the stability of the radially symmetric solutions under symmetry-breaking perturbations. Again, we focus the theoretical discussion in the limiting cases of almost spherical and flat drops, but end with a numerical study of all cases. Finally, section 4 is devoted to the study of the capacity of axially symmetric configurations perturbed in the radial direction and it also includes the proof of several results used in previous sections.

2. Radially symmetric configurations. These are solutions of (1.6) which are invariant under rotations about an axis normal to the solid surface. We can describe the height of each point of the fluid-vapor interface by a function $h(r)$, where r is the radial coordinate in a cylindrical coordinate system about the axis of symmetry. The mean curvature is then (see [7])

$$\kappa = -\frac{1}{r}\frac{d}{dr}\left(r\frac{h_r}{(1+h_r^2)^{\frac{1}{2}}}\right) = -\frac{1}{r}(r\sin\psi)_r, \tag{2.1}$$

where ψ is the angle of inclination of the solution curve $h(r)$ with respect to the r-axis. Notice that $\tan\psi = \frac{dh}{dr}$. We shall study in this section the radially symmetric profiles in two limits for which the analysis simplifies: (1) the limit of small potential V_0 at $\partial\Omega$ so that drops are almost spherical caps, and (2) the limit of large potential V_0 at

$\partial\Omega$ or small contact angle so that drops are almost flat discs. The profiles between these two situations will be found numerically.

2.1. Almost spherical shapes. If we assume $\partial\Omega$ to be continuously differentiable except for the contact line, which we consider located at $r = L$, where $h(r)$ is only Lipschitz continuous (due to the existence of a corner of opening angle θ_Y), then by the classical theory of Dirichlet problems for elliptic linear partial differential equations (cf. [6], for instance) we will have a continuous solution V of problem (1.3)–(1.5) such that $\frac{\partial V}{\partial n}$ will be continuous everywhere except for the contact line. There $\frac{\partial V}{\partial n}$ will be singular since V has the asymptotic behavior

$$V = V_0 + A\rho^{\alpha}\sin(\alpha\theta),$$

where (ρ, θ) are polar coordinates about $(r, z) = (L, 0)$ and α is such that $\sin(\alpha 0) = \sin(\alpha(2\pi - \theta_Y)) = 0$ and $\frac{\partial V}{\partial \rho}$ is square integrable in the neighborhood of the contact line. Hence $\alpha = \frac{1}{2-\theta_Y/\pi}$ and

$$\sigma = -\frac{\partial V}{\partial n} \sim A'\rho^{\frac{1}{2-\theta_Y/\pi}-1} \quad \text{as } \rho \to 0.$$

Notice that σ^2 is integrable with respect to ρ, provided $\theta_Y > 0$, and therefore is integrable over the whole $\partial\Omega$.

Let us assume that $V_0 = \varepsilon \ll 1$ so that we can write the surface charge distribution of a unit volume spherical cap with contact angle θ_Y as $\varepsilon\Sigma$. Then (1.7) and (2.1) yield the equation

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{h_r}{(1+h_r^2)^{\frac{1}{2}}}\right) = -\varepsilon^2\frac{\Sigma^2}{2} + p.$$

By integrating once we get

$$\sin\psi = \frac{1}{2}pr - \varepsilon^2 a_1(r),$$

where

$$a_1(r) = \frac{1}{r}\int_0^r \frac{\Sigma^2}{2}r'dr'.$$

Notice that

$$-\sin\theta_Y = \frac{1}{2}pL - \varepsilon^2 a_1(L). \tag{2.2}$$

In order to find the profile we use

$$\begin{aligned} h_r = \tan\psi &= \frac{\frac{1}{2}pr - \varepsilon^2 a_1(r)}{\sqrt{1-\left(\frac{1}{2}pr - \varepsilon^2 a_1(r)\right)^2}} \\ &= \frac{\frac{1}{2}pr}{\sqrt{1-\frac{1}{4}p^2r^2}} - \frac{1}{\left(1-\frac{1}{4}p^2r^2\right)^{\frac{3}{2}}}\varepsilon^2 a_1(r) + O(\varepsilon^4), \end{aligned}$$

and then

$$h(r) = h_0 + \frac{2}{p} - \frac{2}{p}\sqrt{1-\frac{1}{4}p^2r^2} - \varepsilon^2 a_2(r, p),$$

where

$$a_2(r,p) = \int_0^r \frac{1}{\left(1-\frac{1}{4}p^2r'^2\right)^{\frac{3}{2}}}a_1(r')dr.$$

The parameters h_0, p, and L are chosen so that (2.2) is satisfied, the volume is 1, and $h(L) = 0$. Hence we impose (2.2) together with the condition

$$\begin{aligned} &2\pi\int_0^L \left(h_0 + \frac{2}{p} - \frac{2}{p}\sqrt{1-\frac{1}{4}p^2r^2}\right) rdr - \varepsilon^2 a_3(L,p) \\ &= \pi L^2\left(h_0 + \frac{2}{p}\right) - \frac{16\pi}{3p^3}\left(1-\left(1-\frac{1}{4}p^2L^2\right)^{\frac{3}{2}}\right) - \varepsilon^2 a_3(L,p) = 1, \end{aligned} \tag{2.3}$$

where

$$a_3(r,p) = 2\pi\int_0^r a_2(r',p)r'dr'$$

and

$$0 = h_0 + \frac{2}{p} - \frac{2}{p}\sqrt{1-\frac{1}{4}p^2L^2} - \varepsilon^2 a_2(L,p). \tag{2.4}$$

We write now $p = p_0 + \varepsilon^2 p_1$, $h_0 = h_{0,0} + \varepsilon^2 h_{\prime 0,1}$, $L = L_0 + \varepsilon^2 L_1$ and obtain the following solution at order zero in ε from (2.2), (2.3), (2.4):

$$L_0 = \left[\frac{3}{\pi}\frac{\sin^3\theta_Y}{(\cos\theta_Y - 1)^2(\cos\theta_Y + 2)}\right]^{\frac{1}{3}}, \tag{2.5}$$

$$p_0 = -2\left[\frac{\pi}{3}(\cos\theta_Y - 1)^2(\cos\theta_Y + 2)\right]^{\frac{1}{3}}, \tag{2.6}$$

$$h_{0,0} = \left(\frac{3}{\pi}\frac{1-\cos\theta_Y}{2+\cos\theta_Y}\right)^{\frac{1}{3}}; \tag{2.7}$$

and we obtain the following solutions at $O(\varepsilon^2)$:

$$\begin{aligned} p_1 &= \frac{-a_3p_0^3\cos\theta_Y + 8\pi a_1\sin^3\theta_Y - 8\pi a_2p_0\cos\theta_Y\sin^2\theta_Y}{3p_0^2\cos\theta_Y}, \\ L_1 &= \frac{2}{p_0}\left(-\frac{a_1}{\cos\theta_Y(\cos\theta_Y+2)} + \frac{1}{6}\frac{L_0\left(a_3p_0^3\cos\theta_Y + 8\pi a_2p_0\cos\theta_Y\sin^2\theta_Y\right)}{p_0^2\cos\theta_Y}\right), \\ h_{0,1} &= \frac{2(1-\cos\theta_Y)\left(a_3p_0^3\cos\theta_Y + 8\pi a_2p_0\sin^2\theta_Y\cos\theta_Y\right)}{3p_0^4\cos\theta_Y} + \frac{2a_1\sin\theta_Y}{p_0(\cos\theta_Y+2)} - a_2, \end{aligned}$$

where $a_1 = a_1(L)$, $a_2 = a_2(L,p)$, $a_3 = a_3(L,p)$. Notice that, since $a_i > 0$ for $i = 1,2,3$ and $p_0 < 0$, it follows that $L_1 > 0$ and $h_{0,1} < 0$. Therefore, the drops spread a length $\varepsilon^2 L_1 + O(\varepsilon^4)$.

2.2. Almost flat shapes. In the limit of large values of the potential, or in case $\theta_Y \ll 1$, the drop's configurations are so that the aspect ratio of its height to its radius is a very small number and we can approximate the drop by a flat disc. An advantage to this approximation lies in the fact that the charge distribution in a flat conducting disc has an explicit closed form formula (see [9]) given by

$$\sigma(r) = \frac{Q/(2\pi a)}{\sqrt{a^2 - r^2}}, \tag{2.8}$$

where Q is the total charge stored by the disc and a is its radius. Here we consider the disc to be two-sided and the distribution on each side will be 1/2 of the density $\sigma(r)$ in (2.8). Therefore, (1.7) can be approximated by

$$\kappa - \frac{Q^2/(4\pi a)^2}{2(a^2 - r^2)} = -p,$$

or, in terms of the potential V_0 at the surface of the drop and using the value of the capacity of a flat disc $C = 8a$, by

$$-\frac{1}{r}\frac{d}{dr}\left(r\frac{h_r}{(1+h_r^2)^{\frac{1}{2}}}\right) - \frac{2}{\pi^2}\frac{V_0^2}{(a^2 - r^2)} = -p. \tag{2.9}$$

We introduce now the change of variables in (2.9),

$$p = \frac{V_0^2}{a^2}k, \quad r = ar', \quad h = V_0^2 H,$$

to arrive at

$$\frac{1}{r'}\frac{d}{dr'}\left(r'\frac{H_{r'}}{(1+\frac{V_0^2}{a^2}H_{r'}^2)^{\frac{1}{2}}}\right) + \frac{2}{\pi^2}\frac{1}{(1 - r'^2)} = k$$

and assume $\frac{V_0^2}{a^2} \ll 1$ so that we can finally deduce the equation

$$\frac{1}{r'}\frac{d}{dr'}(r'H_{r'}) + \frac{2}{\pi^2}\frac{1}{(1 - r'^2)} = k,$$

with solution

$$H(r') = -\frac{1}{4}k + \frac{1}{12} + \frac{1}{4\pi^2}(4\,\mathrm{dilog}(1 - r') + 4\,\mathrm{dilog}(1 + r') + \pi^2 k r'^2), \tag{2.10}$$

such that $H(1) = 0$. A few profiles $H(r')$ for various values of k are represented in Figure 2.1. The fact that the drop's center touches the solid substrate when $k = \frac{1}{3}$, as one can easily compute from (2.10), implies that these solutions cannot be constructed for arbitrary values of a and V. If we compute the volume associated to these drops, we find

$$2\pi\int_0^1 H(r')r'dr' = \frac{Vol.}{V_0^2 a^2},$$

or, using (2.10),

$$\frac{0.6366}{4}\cdots - \frac{1}{8}\pi k = \frac{Vol.}{V_0^2 a^2},$$

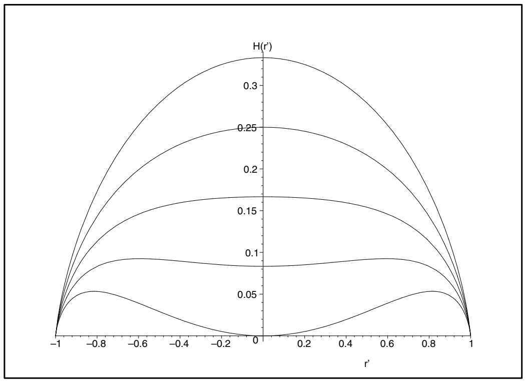


FIG. 2.1. *Profiles $H(r')$ for $k = 0, 1/8, 1/6, 1/4, 1/3$.*

implying, since $k < \frac{1}{3}$,

$$V_0^2 a^2/4 < 8.847\ldots. \tag{2.11}$$

The immediate implication of this formula is that one cannot spread the drop indefinitely even when it is strongly charged.

2.3. The calculation of shapes for moderate values of $\boldsymbol{V_0}$ and $\boldsymbol{\theta_Y}$. We have implemented a numerical method in order to calculate stationary shapes of drops. For a given value of the potential V_0 we find the shapes for varying values of the base radius a and compute their energy. This will allow us to compute the solid-liquid interface radius of the equilibrium shapes for given values V_0 and θ_Y by minimization of the energy (1.8) as a function of a. The way to approach the problem is by finding solutions of (1.7) as stationary solutions of the evolution problem

$$h_t - \Delta\left(\kappa - \frac{\sigma^2}{2}\right) = 0, \tag{2.12}$$

with boundary conditions

$$h(a) = \left.\frac{\partial(\kappa - \frac{\sigma^2}{2})}{\partial n}\right|_{r=a} = 0. \tag{2.13}$$

As the initial condition we take a spherical cap with base radius a. We compute numerically the solutions to problem (2.12), (2.13) and observe that $\kappa - \frac{\sigma^2}{2}$ converges to a constant value $-p$ and stop the calculation when the difference between the maximum and minimum values of $\kappa - \frac{\sigma^2}{2}$ lies below some tolerance threshold fixed a priori. At each time-step we compute σ by solving the integral equation

$$V_0 = \frac{1}{4\pi}\int_{\partial\Omega} \frac{\sigma(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dS \tag{2.14}$$

with a boundary element technique for axisymmetric profiles (cf. [16]).

We compute, for these profiles, the energy as the sum of two contributions,

$$\begin{aligned} E &= E_1 - \cos\theta_Y E_2, \\ E_1 &= A_{lv} - \frac{1}{2}CV_0^2, \quad E_2 = A_{sl} = \pi a^2. \end{aligned}$$

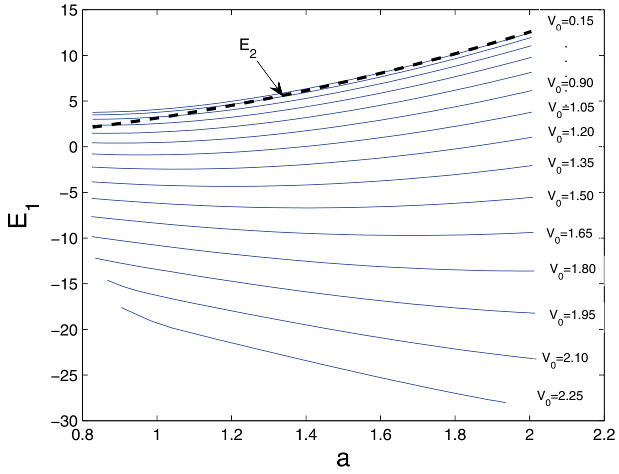


FIG. 2.2. *Energies as a function of the radius of the circular base* (a) *for different potentials* V_0. *The potentials run from* 0.15 *up to* 2.25 *at intervals of* 0.15.

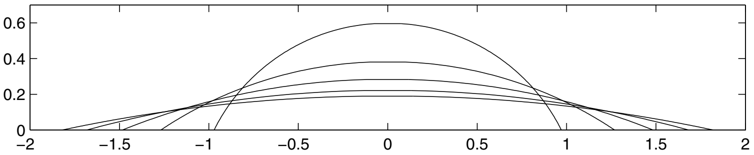


FIG. 2.3. *Shapes of the profiles for various values of the radius of the circular base* (a) *and* $V_0 = 0.15$.

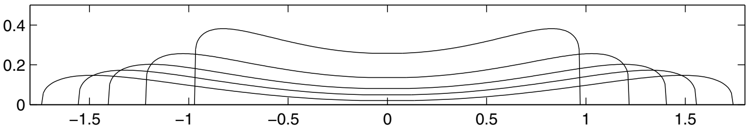


FIG. 2.4. *Shapes of the profiles for various values of the radius of the circular base* (a) *and* $V_0 = 2.25$.

In Figure 2.2 we represent the computed values of E_1 for various values of V_0. The profiles for $V_0 = 0.15$ and 2.25 are represented in Figures 2.3 and 2.4, respectively. For a given value of θ_Y one can compute the equilibrium shapes for each value of V_0 by minimizing E as a function of a, that is, by finding the solutions to

$$\frac{dE}{da} = \frac{dE_1}{da} - \cos\theta_Y \frac{dE_2}{da} = 0,$$

leading to the relation

$$\theta_Y = \arccos\left(\frac{\frac{dE_1}{da}}{\frac{dE_2}{da}}\right).$$

In Figure 2.5 we represent θ_Y as a function of a for various V_0. The intersection of the horizontal line corresponding to a given θ_Y with the curves of constant V_0 gives the

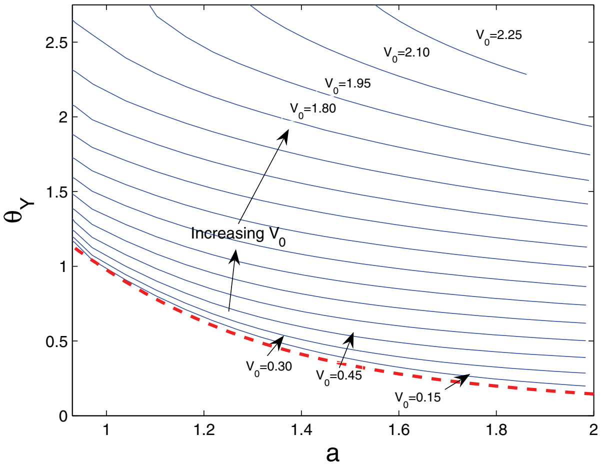


FIG. 2.5. *θ_Y as a function of the radius of the circular base (a) for different values of the potential V_0. The potentials run from 0.15 up to 2.25 at intervals of 0.15. The dashed line is the explicit curve, at $V_0 = 0$, of L_0 $(= a)$ as a function of θ_Y given by* (2.5).

values of a as a function of V_0. Notice that we are representing level lines of constant V_0 at intervals of 0.15 and that the level lines increase their separation with increasing V_0. This implies a superlinear growth of a with V_0 for a given value of θ_Y. As we saw in section 2.1, this increase is exactly quadratic for sufficiently small V_0.

2.4. The approximation with ellipsoids. When $V_0 = 0$, the axially symmetric minimizers of (1.8) are spherical caps. In the case of levitating droplets ($\theta_Y = \pi$) the minimizers are spheres if $V_0 < 1.575\ldots$ and approximately oblate ellipsoids (cf. [8], [4], and section 3.1 below) if $V_0 > 1.575\ldots$. Moreover, we have observed that for moderate values of the potential the profiles can be approximated by truncated oblate ellipsoids (ellipsoidal caps) with a very high degree of accuracy. These observations lead us to study the variational problem (1.8) restricted to the class of ellipsoidal caps in the neighborhood of $\theta_Y = \pi$. We shall see below that the resulting $a - \theta_Y$ diagram is consistent with Figure 2.5 and that both overlap smoothly in the common region.

The family of profiles we consider is

$$\frac{1}{c^2}(z - \alpha c)^2 + \beta^2 r^2 = 1$$

for $z \geq 0$ only. These are ellipsoidal caps resulting from the truncation of an axially symmetric ellipsoid with the center at αc and semiaxis c and β^{-1}. Written in this form, it is simple to compute c such that the volume of the drop is 1. Hence we have to consider a two-parameter family of profiles with parameters $-1 \leq \alpha \leq 1$ and $\beta \leq \frac{1}{c}$. The areas A_{lv} and A_{sl} and the capacity C are then functions of α and β, and the minimization problem is simply the one of finding the minima of a function of two variables. We can show that $A_{sl} = \pi \frac{1-\alpha^2}{\beta^2}$. The main difficulty is the lack of explicit expressions (except for just a few cases) for A_{lv} and C. We compute these values numerically for (α, β) in a grid with $O(10^5)$ nodes. The capacity is determined as the total charge from a surface charge distribution σ that we evaluate by solving the integral equation (2.14) with a boundary element method. Then, for given values of

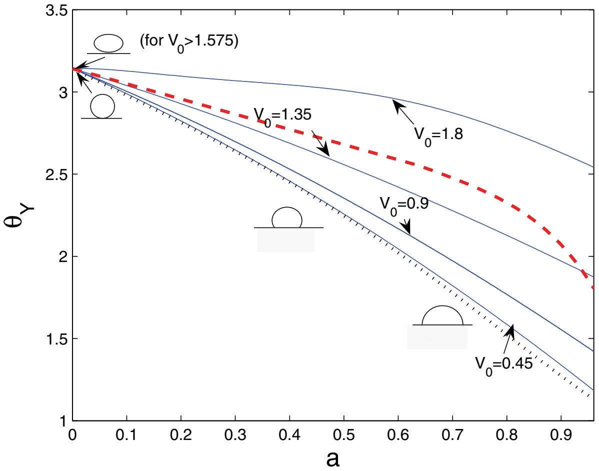


Fig. 2.6. *θ_Y as a function of the radius of the circular base (a) for different values of the potential V_0 using the approximation with truncated ellipsoids. The dotted line is the explicit curve at $V_0 = 0$. The dashed line represents the line of symmetry breaking bifurcations to be discussed in section* 3.

V_0, we compute for each θ_Y the values of α and β in the grid that minimize the energy (1.8), and with these values we compute a. The result is represented in Figure 2.6. In the range of values of a and θ_Y for which Figures 2.5 and 2.6 overlap, the ellipsoidal approximation is very accurate. Nevertheless, it degrades for larger values of a, and one can only apply it safely in the parametric region of Figure 2.6.

3. Instabilities of radially symmetric solutions. In this section we shall proceed to study the stability of the radially symmetric solutions deduced above. We approach the problem in a spirit very similar to Rayleigh's in [17], that is, by computing the energy associated to charged axially symmetric droplets and looking at situations where the breaking of axial symmetry via a small perturbation is more favorable energetically. The existence of these energetically preferred nonsymmetric configurations in the neighborhood of a symmetric configuration indicates that in a dynamic situation, such as, for instance, a gradient flow associated to this energy or a thin-film-type equation [2], symmetric solutions might become unstable. Once a symmetric solution destabilizes there are essentially two possibilities: evolving towards a different equilibrium configuration with broken symmetry, or evolving in time up to the formation of some kind of singularity in the solution of the system of evolution partial differential equations (such as Navier–Stokes for fluids). We are not going to study the evolution problem here, but our discussion will show when instabilities happen and which kind of perturbation modes make the energy decrease most so that one can expect them to be the dominant ones in the early-time evolution of the system and somehow determine the shape of the asymmetric drop. As in the previous section, we start with two limiting cases: the spherical drops and the flat drops, where analytical calculations are feasible, and study all other cases numerically.

3.1. Almost spherical shapes. When $\theta_Y = \pi$, there are equilibrium configurations which are perfectly spherical independent of the potential at their boundary V_0. These configurations are certainly stable when V_0 is sufficiently small, but eventually will become unstable if they store a large charge (and have a large V_0 consequently). In what follows, we will compute the energy of nearby spheroidal configurations and deduce a formula for the value of V_0 at which spheroidal shapes are preferred to spherical shapes. This analysis can easily be extended to drops with different, but close to π, values of θ_Y.

For an oblate ellipsoid with semiaxis a and c such that $c < a$, the volume, area, and capacity (see [14] and [15]) are given by

$$\begin{aligned} Vol. &= \frac{4}{3}\pi a^2 c, \\ A &= 2\pi \left(a^2 + c^2 \frac{\operatorname{arctanh}(\sin\alpha)}{\sin\alpha} \right), \\ C &= 4\pi \frac{a\beta}{\arcsin\beta}, \end{aligned}$$

where

$$\alpha = \arccos(c/a), \tag{3.1a}$$

$$\beta = \sqrt{1 - \frac{c^2}{a^2}}. \tag{3.1b}$$

For a prolate ellipsoid with semiaxis a and c such that $a < c$,

$$\begin{aligned} Vol. &= \frac{4}{3}\pi a^2 c, \\ A &= 2\pi \left(a^2 + c^2 \frac{\alpha}{\tan\alpha} \right), \\ C &= 8\pi \frac{a\beta}{\log\frac{1+\beta}{1-\beta}}, \end{aligned}$$

with α and β given by (3.1a), (3.1b). Hence, the energy associated to both oblate and prolate spheroids is, respectively,

$$\begin{aligned} E_{oblate} &= 2\pi \left(a^2 + c^2 \frac{\operatorname{arctanh}(\sin\alpha)}{\sin\alpha} \right) - \frac{1}{2}\left(4\pi \frac{a\beta}{\arcsin\beta} \right) V_0^2, \\ E_{prolate} &= 2\pi \left(a^2 + c^2 \frac{\alpha}{\tan\alpha} \right) - \frac{1}{2}\left(8\pi \frac{a\beta}{\log\frac{1+\beta}{1-\beta}} \right) V_0^2. \end{aligned}$$

If $a = c = R$, the case of a sphere, then

$$E_{oblate} = E_{prolate} = 4\pi R^2 - 2\pi R V_0^2.$$

If we perturb the sphere to be an ellipsoid (with the same volume) by considering

$$\begin{aligned} a &= R(1+\varepsilon), \quad c = \frac{R}{(1+\varepsilon)^2} \quad \text{in the oblate case,} \\ a &= \frac{R}{(1+\varepsilon)}, \quad c = R(1+\varepsilon)^2 \quad \text{in the prolate case} \end{aligned}$$

($0 < \varepsilon \ll 1$), we get the following energies:

$$\begin{aligned} E_{oblate} &= (4\pi R^2 - 2\pi R V_0^2) + \left(-\frac{8}{5}\pi R V_0^2 + \frac{32}{5}\pi R^2 \right)\varepsilon^2 + O(\varepsilon^3), \\ E_{prolate} &= (4\pi R^2 - 2\pi R V_0^2) + (8\pi R^2 - 2\pi R V_0^2)\varepsilon + O(\varepsilon^2). \end{aligned}$$

In both types of perturbations, the ellipsoidal shapes are more favorable energetically when

$$V_0^2 > 4R. \tag{3.2}$$

In fact, since $\varepsilon \ll 1$, the prolate shape should be more favorable. A spherical drop such that (3.2) is satisfied shall become unstable, and preferably start its evolution by deforming into a prolate spheroid. This is, in fact, what is observed in experiments done on levitating droplets [5] and in the numerical simulations in [3].

For a sphere of unit volume, (3.2) holds when

$$V_0 > 2\left(\frac{3}{4\pi}\right)^{\frac{1}{6}} = 1.575\ldots. \tag{3.3}$$

3.2. Almost flat shapes. We consider now very extended drops such that the radius of their base is $a \gg 1$ and their height $h_0 = h(0) \ll 1$. When $V_0 = 0$ these drops are spherical caps and they appear when $\theta_Y \ll 1$. If the volume, given by

$$Vol. = \frac{1}{6}\pi h_0(3a^2 + h_0^2), \tag{3.4}$$

is equal to 1, then it follows from a direct calculation that, for $V_0 = 0$, a and h_0 are given by

$$h_0 = \left(\frac{3}{\pi}\frac{1-\cos\theta_Y}{2+\cos\theta_Y}\right)^{\frac{1}{3}}, \quad a = \left(\frac{3}{\pi}\right)^{\frac{1}{3}}\left(\frac{1-\cos\theta_Y}{2+\cos\theta_Y}\right)^{-\frac{1}{6}}\left(\frac{1+\cos\theta_Y}{2+\cos\theta_Y}\right)^{\frac{1}{2}}, \tag{3.5}$$

which coincide with the expression we obtained for $h_{0,0}$ and L_0 in (2.7) and (2.5), respectively.

The capacity of a flat disc is $C = 8a$ and the area of a spherical cap is $\pi(a^2 + h_0^2)$. Hence, the total energy of the charged drop is

$$E = \pi(a^2 + h_0^2) - (\cos\theta_Y)\,\pi a^2 - \frac{1}{2}CV_0^2 = \pi(a^2 + h_0^2) - (\cos\theta_Y)\,\pi a^2 - V_0^2 4a,$$

or keeping in mind the relation between a and h_0, from (3.4) and the constraint $Vol. = 1$,

$$E(h_0) = \pi\left(\frac{2}{\pi h_0} - \frac{h_0^2}{3} + h_0^2\right) - (\cos\theta_Y)\,\pi\left(\frac{2}{\pi h_0} - \frac{h_0^2}{3}\right) - V_0^2 4\sqrt{\frac{2}{\pi h_0} - \frac{h_0^2}{3}}.$$

For small h_0,

$$E'(h_0) \simeq (1-\cos\theta_Y)\left(\frac{-2}{h_0^2}\right) + 2\frac{V_0^2}{h_0^{\frac{3}{2}}}\sqrt{\frac{2}{\pi}} + O(h_0) = 0,$$

and hence

$$h_0 \sim \frac{\pi}{2V_0^4}\left(1-\cos\theta_Y\right)^2, \tag{3.6}$$

$$a = \frac{1}{h_0}\left(\frac{2}{\pi} - \frac{h_0^3}{3}\right) \sim \frac{4V_0^4}{\pi^2\left(1-\cos\theta_Y\right)^2}, \tag{3.7}$$

which gives an estimate of the radius of the drop's base as a function of V_0 and θ_Y, provided V_0 is small enough for (2.11) to be verified and θ_Y is sufficiently close to zero in order to have almost flat drops.

Next we will discuss whether or not almost flat configurations which are not discs can be more favorable energetically than the axially symmetric configurations. This is not surprising due to the important result of Pólya and Szegö (cf. [14]) on the effect of Schwarz symmetrization on capacity. They proved that the process of Schwarz symmetrization of a three-dimensional body diminishes (or leaves unchanged) the capacity of a body. Schwarz symmetrization about an axis e replaces each cross section of the body orthogonal to e by a circular section of the same area centered at e. This leads to a body of revolution. In the case of a disc, one can conclude that the capacity of a perturbed disc is larger than or equal to the capacity of the disc of the same area. Therefore, the energy functional decreases or remains equal when deforming a disc. We will show below that the energy, in fact, decreases.

The energy of the almost flat droplet can be approximated by

$$E \sim (1 - \cos\theta_Y)\, A_{sl} - \frac{1}{2} C_b V_0^2,$$

where we have assumed $h_r(r) \ll 1$ so that one can assume that the liquid-vapor interface has an area similar to the liquid-solid interface and the capacity of the drop is similar to the capacity of the planar region consisting of the liquid-solid interface, denoted by C_b.

When changing the geometry of the liquid-solid interface, the energy will experience a variation given by

$$\delta E = (1 - \cos\theta_Y)\, \delta A_{sl} - \frac{1}{2} V_0^2 \delta C_b.$$

A flat ellipsoid has capacity (cf. [14]) and area given by

$$C_b = 2\pi \frac{b' + b}{K\left(\frac{b'-b}{b'+b}\right)}, \quad A_{sl} = \pi b' b,$$

where b and b' are the length of the semiaxis of the ellipse and $K(x)$ is the complete elliptic integral of the first kind. If we perturb the disc into an ellipse, in such a way that the area is preserved, by writing

$$b = a(1+\varepsilon), \quad b' = \frac{a}{1+\varepsilon},$$

we will have

$$\delta A_{sl} = 0, \quad \delta C_b = \frac{2a + a\varepsilon^2}{K\left(\varepsilon - \frac{1}{2}\varepsilon^2\right)} - 8a2\pi \sim 2\pi \frac{2a + a\varepsilon^2}{\frac{\pi}{2} + \frac{\pi}{8}\varepsilon^2} - 8a \sim 8a\varepsilon^2,$$

and this implies $\delta E < 0$. Therefore, the flat ellipsoid is energetically favorable with respect to a flat disc. The same result will hold true for nonplanar bodies which are almost flat in the sense that h_0 given by (3.6) is much smaller than a, given by (3.7).

Notice that (3.7) has to be combined with (2.11) in order to have consistent almost-flat droplets. This condition requires, in addition to a given by (3.7) being large, that

$$V_0 a = \frac{4V_0^5}{\pi^2 \left(1 - \cos\theta_Y\right)^2} < 2\sqrt{8.848\ldots},$$

which is only true if V_0 is sufficiently small.

Finally, we introduce general perturbations of the flat disc in the form

$$r(\theta) = a\frac{1}{1+\frac{\varepsilon^2}{2}}(1+\varepsilon\cos(n\theta)),$$

with θ and r polar coordinates about the axis, so that

$$A_{sl} = \frac{1}{2}\int r^2(\theta)d\theta = a^2. \tag{3.8}$$

When $n = 2$, we recover, at first order in ε, the perturbation into an ellipsoid. For $n > 2$ we obtain a shape with n undulations of the disc's boundary. Such perturbations do not change the area by (3.8) but do change the capacity in the sense of increasing it with increasing n, provided $n\varepsilon \ll 1$. This implies that the larger n is, the more favorable energetically is a configuration with n undulations in its boundary.

3.3. Symmetry-breaking instabilities for intermediate values of $\boldsymbol{V_0}$ and $\boldsymbol{\theta_Y}$. We identify values of the potential at which nonaxisymmetric solutions are energetically more favorable than the axially symmetric configurations $r = a(z)$. We perturb the profiles radially in the form

$$r(\theta, z) = \frac{a(z)}{\sqrt{1+\frac{\varepsilon^2}{2}}}(1+\varepsilon\cos(n\theta)); \quad z\in[0,H], \quad \theta\in[0,2\pi). \tag{3.9}$$

In the last section, we will show that this kind of perturbation leads, for $\varepsilon \ll 1$ and provided $n\varepsilon \ll 1$, to an increase of the capacity which is a nondecreasing function of n. Since this fact follows from a lengthy calculation we postpone its proof and use it here to conclude the existence of symmetry-breaking bifurcations.

The volume of the drop does not change, since

$$Vol. = \int_0^H z\left[\int_0^{2\pi}\left(\int_0^{\frac{a(z)}{\sqrt{1+\frac{\varepsilon^2}{2}}}(1+\varepsilon\cos(n\theta))} rdr\right)d\theta\right]dz$$

$$= \int_0^H z\left[\frac{1}{2}\int_0^{2\pi}\frac{a^2(z)}{1+\frac{\varepsilon^2}{2}}(1+\varepsilon\cos(n\theta))^2 d\theta\right]dz = \pi\int_0^H za^2(z)dz.$$

The lateral area (area of the liquid-vapor interface) has a value

$$\begin{aligned}
A_{lv}(\varepsilon) &= \int_0^H\int_0^{2\pi}\sqrt{r^2+r_\theta^2+r^2r_z^2}d\theta dz \\
&= \int_0^H\int_0^{2\pi}\left(a^2\left(1+2\varepsilon\cos n\theta+\frac{\varepsilon^2}{2}\cos 2n\theta\right)+a^2n^2\varepsilon^2\sin^2(n\theta)\right. \\
&\qquad \left. +a^2a_z^2(1+4\varepsilon\cos n\theta+\varepsilon^2(3\cos 2n\theta+2))\right)^{\frac{1}{2}}dz+O(\varepsilon^3) \\
&= A_{lv}(0)+\varepsilon^2\pi\int_0^H\frac{\frac{n^2}{2}+2a_z^2}{\sqrt{1+a_z^2}}adz+O(\varepsilon^3) \\
&= A_{lv}(0)+\varepsilon^2(c_1n^2+c_2)+O(\varepsilon^3),
\end{aligned}$$

while the area of the solid-liquid interface remains unchanged. Hence, we can estimate the variation of the energy under a radial perturbation:

$$\begin{aligned}\delta E &= \delta A_{lv} - \cos\theta_Y \delta A_{sl} - \frac{1}{2}\delta C V_0^2\\ &= \varepsilon^2\left[(c_1 n^2 + c_2) - \frac{1}{2}\left(\delta C/\varepsilon^2\right) V_0^2\right] + O(\varepsilon^3).\end{aligned}$$

Since $\delta C/\varepsilon^2$ is always positive and $O(1)$ for $n > 2$, as we prove in the next section, one should always have, for

$$V_0 > V_{0,n} = \sqrt{\frac{2(c_1 n^2 + c_2)}{(\delta C/\varepsilon^2)}},$$

instabilities with a given n. For profiles close to a sphere $V_{0,n}$ is increasing with n, but this is not the case for profiles close to a disc. Given the strict monotonicity of the capacity with n (even for profiles that are close to discs), again as we prove in the next section, and given the fact that c_1 and c_2 can be made arbitrarily small for sufficiently flat profiles (when $H \ll 1$), we can make the combination $\left[(c_1 n^2 + c_2) - \frac{1}{2}\left(\delta C/\varepsilon^2\right) V_0^2\right]$ to decrease monotonically with n for configurations close to a disc. This would imply that perturbations with large n are more unstable than perturbations with a small n. Hence, one can expect the boundary to destabilize first with multiple oscillations. In the next subsection we implement a numerical method to compute the energy of radially perturbed equilibrium shapes and determine, for given θ_Y, the values of the radius of the solid-liquid interface and V_0 for which instabilities with various n appear.

3.4. Numerical results.

Numerical approximation in the radially perturbed case. When dealing with radially perturbed equilibrium shapes we lose the axially symmetric properties and need to do a full three-dimensional (3D) approximation in order to compute area and capacity and hence the energy given in (1.8). We use a boundary element method that we have already implemented in [3] to compute the surface charge density by solving the integral equation (2.14) in the profiles given by (3.9) with a unit potential at the boundary. From the surface charge density we can obtain the total charge integrating over the surface and that will be the capacity of the body.

Validation. In order to validate the numerical method described above we have done two tests:

1. Comparison with the exact capacity of a perfectly conducting thin spherical shell. The exact capacity can be obtained through

$$C = 4a(\sin\beta + \beta),$$

where a is the radius of the sphere and β is the zenithal angle of the shell; see [1].

We have compared in Figure 3.1 the capacity obtained numerically for values of β from 0 up to $\pi/2$ (a hemispherical shell) and with two different meshes (mesh 1 and mesh 2; see Figure 3.2), with the exact capacity. We found with mesh 1 a maximum relative error of 0.011 and with mesh 2 a maximum relative error of 0.0058. In both cases, the relative error is below 2%.

2. Comparison with the exact capacity of a hemisphere (a spherical cap with $\beta = \pi/2$ and a lower tap representing the liquid-solid interface). The exact capacity

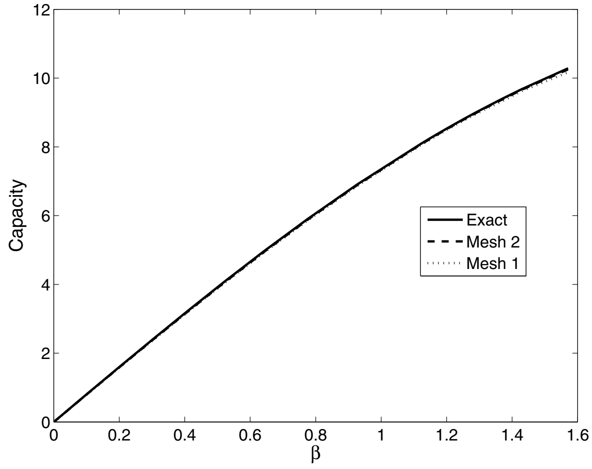


FIG. 3.1. *Capacity versus zenithal angle in an axisymmetric drop.*

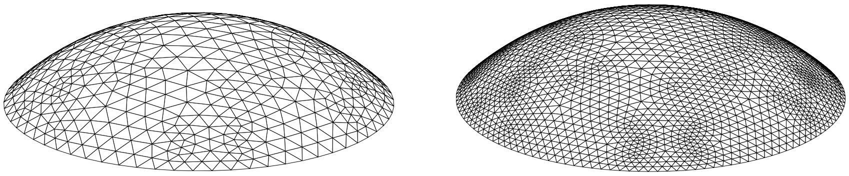


FIG. 3.2. *Two* 3*D meshes of a spherical cap (*$a = 0.7$*,* $\beta = 0.7754$*). On the left, mesh* 1 *(*541 *nodes and* 1016 *elements), and on the right, mesh* 2 *(*2097 *nodes and* 4064 *elements).*

of a hemisphere of radius a is (see [10])

$$4\pi a\left(1-\frac{1}{\sqrt{3}}\right).$$

In this case we compare again the results obtained with meshes 1 and 2. We obtained a relative error of 0.00997 with mesh 1 and a relative error of 0.00253 with mesh 2.

Results and discussion. We have used the numerical method described above to compute the energy of profiles perturbed in the form (3.9) and compare it with the energy of the axisymmetric profile $a(z)$. We use mesh 2 for the numerical simulations. In Figure 3.3 we show the perturbed drops with $n = 0$, 2, 3, and 4 for the case corresponding to $V = 0.6$ and $a = 1.16$.

In Figure 3.4 we represent, together with the curves of θ_Y as a function of a for all axisymmetric profiles at a given potential, the approximate "bifurcation curves" for $n = 2$, 3, and 4. Each of these curves, with a given value of n, delimits the regions where the configurations of the type (3.9) with $\varepsilon \ll 1$ (that we took equal to 0.025 for our numerical computations) are energetically more favorable than configurations with smaller value of n (including the axisymmetric profiles). Notice the tendency of the curves to intersect for large values of a. This is in agreement with the discussion on profiles close to a disc in the previous section.

From Figure 3.4 we can obtain two important conclusions:

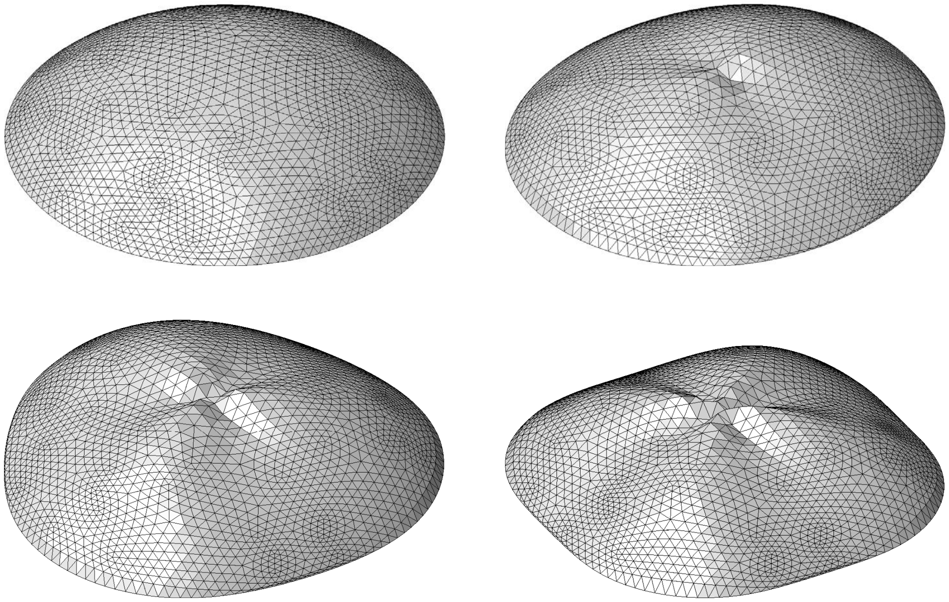


Fig. 3.3. *Shape of a radially symmetric droplet with a bottom tap of radius $a = 1.16$ under a potential $V = 0.6$ (top left). The same drop perturbed radially with $n = 2$ (top right), $n = 3$ (bottom left), and $n = 4$ (bottom right).*

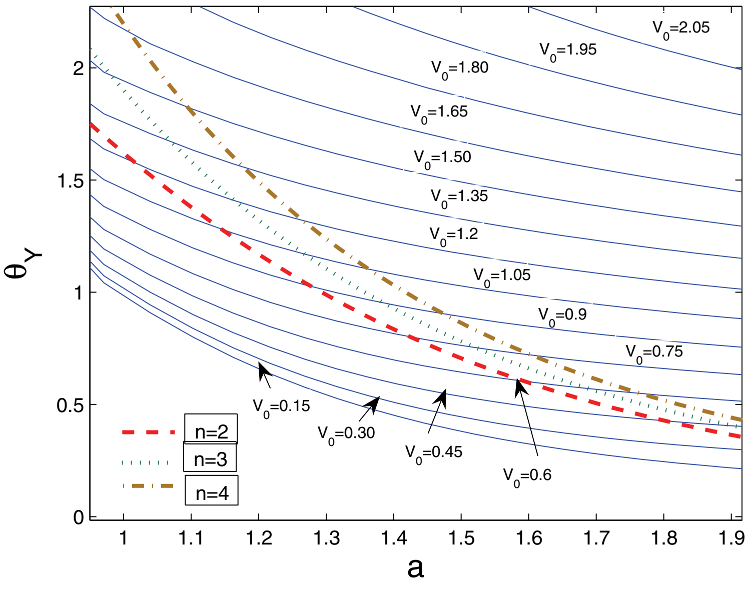


Fig. 3.4. *Bifurcation branches in the $a - \theta_Y$ plane. The bifurcations with $n = 2$ are represented with dashed line, $n = 3$ with dotted line, and $n = 4$ with dash-dot line.*

1. Drops cannot be spread indefinitely by increasing the potential V_0. If we trace a horizontal line in Figure 3.4 for a given value of θ_Y, we find that a increases with the potential in a superlinear manner, but only up to some limiting value where

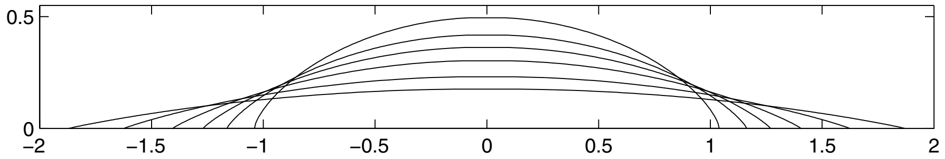


FIG. 3.5. *Axisymmetric profiles at the bifurcation curve $n = 2$ for $\theta_Y = 1.55$, 1.26, 1.03, 0.81, 0.59, 0.43.*

we meet the first bifurcation curve $n = 2$. At this point, an axisymmetric drop is no longer the most favorable configuration, and drops with elliptic cross sections will be preferred. A further increase of V_0 leads to an intersection with the curves $n = 3, 4, \ldots$ and configurations with a higher number of undulations would be more favorable energetically. The intersection with the first bifurcation curve takes place for relatively small values of a (in comparison with the maximum values that a may take from the constraint established in (2.11)). In Figure 3.5 we represent the axisymmetric profiles at the $n = 2$ bifurcation curve for different values of θ_Y. Observe that all the profiles are concave. Hence, the profiles changing concavity, like those represented in Figure 2.4, are probably not observable in nature since nonaxisymmetric perturbations lead to configurations which are energetically more favorable. Once a bifurcation curve has been crossed, axisymmetric drops should destabilize. This provides an explanation to the saturation effect in electrowetting.

2. For a given value of θ_Y, the transition between bifurcation curves occurs relatively fast: if we take, for instance, $\theta_Y = \frac{\pi}{4}$, we can see that a only changes approximately from 1.4 to 1.6 while crossing the bifurcation curves $n = 2$, 3, and 4. This fast transition is more remarkable the smaller θ_Y is, and for sufficiently small values of θ_Y the transition takes place for very flat shapes; for those shapes (see the previous section) the first bifurcation curve we cross when increasing a may correspond to $n > 2$. The quick transition between bifurcation curves provides an explanation of the characteristic multiple-finger patterns observed in electrowetting once saturation is reached and potential is slowly increased.

4. The capacity of radially perturbed bodies of revolution. In this section we deduce some important results used in this paper concerning the capacity of a radially perturbed body of revolution (including drops as a particular case). We shall show how the breaking of axial symmetry leads to an increase in the capacity, and we estimate this increase. By Dirichlet's principle, the capacity of Ω can be evaluated as (cf. [14])

$$C = \int_{\mathbb{R}^3 \setminus \Omega_0} |\nabla V|^2 \, d\mathbf{x}.$$

We will consider here the case of Ω being a small axial perturbation of an axisymmetric domain Ω_0. More precisely, we shall consider a domain described in cylindrical coordinates (r, θ, z) by

$$r(\theta, z) = \frac{a(z)}{\sqrt{1 + \frac{\varepsilon^2}{2}}}(1 + \varepsilon \cos(n\theta)); \quad z \in [0, H], \quad \theta \in [0, 2\pi),$$

where $r = a(z)$ describes the generatrix of Ω_0; together with the upper and lower taps $r < \{a(H)(1 + \varepsilon \cos(n\theta)),\, z = H\}$ and $r < \{a(0)(1 + \varepsilon \cos(n\theta)),\, z = 0\}$. Notice that

$a(H) = 0$ in the particular case of a drop, which is the one considered in the present paper. Then, assuming $a(z) = 0$ for $z < 0$ and $z > H$, one can write

$$\int_{\mathbb{R}^3 \setminus \Omega_0} |\nabla V|^2 \, d\mathbf{x} = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_{\frac{a(z)}{\sqrt{1+\frac{\varepsilon^2}{2}}}(1+\varepsilon\cos(n\theta))}^{\infty} |\nabla V|^2 \, r dr d\theta dz.$$

We change variables to a new system $r' = \sqrt{1+\frac{\varepsilon^2}{2}}\, r/(1+\varepsilon\cos(n\theta))$, $\theta' = \theta$, $z' = z$, so that

$$\frac{\partial V}{\partial r} = \frac{\partial V}{\partial r'} \frac{1}{1+\varepsilon\cos(n\theta')},$$
$$\frac{\partial V}{\partial \theta} = \frac{\partial V}{\partial \theta'} + \frac{\partial V}{\partial r'} r' \frac{\varepsilon n \sin(n\theta')}{1+\varepsilon\cos(n\theta')},$$

and hence

$$\begin{aligned} C = \int_{\mathbb{R}^3 \setminus \Omega_0} \Bigg\{ & \frac{1}{(1+\varepsilon\cos(n\theta'))^2} \left| \frac{\partial V}{\partial r'} \right|^2 + \frac{1}{r'^2 (1+\varepsilon\cos(n\theta'))^2} \left| \frac{\partial V}{\partial \theta'} + \frac{\partial V}{\partial r'} \frac{r' \varepsilon n \sin(n\theta')}{1+\varepsilon\cos(n\theta')} \right|^2 \\ & + \frac{1}{1+\frac{\varepsilon^2}{2}} \left| \frac{\partial V}{\partial z'} \right|^2 \Bigg\} (1+\varepsilon\cos(n\theta'))^2 \, r' dr' d\theta' dz' \\ = \int_{\mathbb{R}^3 \setminus \Omega_0} & \left\{ \left| \frac{\partial V}{\partial r'} \right|^2 + \frac{1}{r'^2} \left| \frac{\partial V}{\partial \theta'} \right|^2 + \left| \frac{\partial V}{\partial z'} \right|^2 \right\} r' dr' d\theta' dz' \\ + \int_{\mathbb{R}^3 \setminus \Omega_0} & \Bigg\{ \frac{2}{r'} \frac{\partial V}{\partial \theta'} \frac{\partial V}{\partial r'} \frac{\varepsilon n \sin(n\theta')}{(1+\varepsilon\cos(n\theta'))} + \frac{1}{r'^2} \left| \frac{\partial V}{\partial r'} r' \frac{\varepsilon n \sin(n\theta')}{1+\varepsilon\cos(n\theta')} \right|^2 \\ & + \left(2\varepsilon\cos(n\theta') + \varepsilon^2 \cos^2(n\theta') - \frac{\varepsilon^2}{2} \right) \left| \frac{\partial V}{\partial z'} \right|^2 \Bigg\} r' dr' d\theta' dz' + O(\varepsilon^3). \end{aligned}$$

Let $V_0(r,z)$ be the potential outside Ω_0 such that $V_0 = 1$ at $\partial\Omega_0$, with C_0 the capacity of Ω_0, and write

$$V(r', \theta', z') = V_0(r', z') + \varepsilon^2 V_1(r', \theta', z').$$

Then, as long as $n\varepsilon \ll 1$, we can expand

$$\begin{aligned} C = & \int_{\mathbb{R}^3 \setminus \Omega_0} |\nabla V_0|^2 \, d\mathbf{x}' + 2\varepsilon \int_{\mathbb{R}^3 \setminus \Omega_0} \nabla V_0 \cdot \nabla V_1 d\mathbf{x}' + \varepsilon^2 \int_{\mathbb{R}^3 \setminus \Omega_0} |\nabla V_1|^2 \, d\mathbf{x}' \\ & + \varepsilon^2 \int_{\mathbb{R}^3 \setminus \Omega_0} \Bigg\{ \frac{2}{r'} \frac{\partial V_1}{\partial \theta'} \frac{\partial V_0}{\partial r'} n \sin(n\theta') + \left(\frac{\partial V_0}{\partial r'} \right)^2 n^2 \sin^2(n\theta') \\ & \qquad + \frac{1}{2} \cos(2n\theta') \left(\frac{\partial V_0}{\partial z'} \right)^2 + 4\cos(n\theta') \frac{\partial V_1}{\partial z'} \frac{\partial V_0}{\partial z'} \Bigg\} r' dr' d\theta' dz' + O(\varepsilon^3), \end{aligned}$$

and since $\int_{\mathbb{R}^3 \setminus \Omega_0} \nabla V_0 \cdot \nabla V_1 d\mathbf{x}' = 0$ (as one can show after integration by parts and using $\Delta V_0 = 0$ outside Ω_0 and $V_1 = 0$ at $\partial\Omega_0$) we can write

$$C = C_0 + \varepsilon^2 C_{n,1} + O(\varepsilon^3),$$

where

$$C_{n,1} = \int_{\mathbb{R}^3 \diagdown \Omega_0} |\nabla V_1|^2 \, d\mathbf{x}' + \int_{\mathbb{R}^3 \diagdown \Omega_0} \Bigg\{ \left(\frac{\partial V_0}{\partial r'} \right)^2 n^2 \sin^2(n\theta') \\ + \frac{2}{r'} \frac{\partial V_1}{\partial \theta'} \frac{\partial V_0}{\partial r'} n \sin(n\theta') + 4 \cos(n\theta') \frac{\partial V_1}{\partial z'} \frac{\partial V_0}{\partial z'} \Bigg\} r' dr' d\theta' dz'.$$

By Dirichlet's principle, V_1 can also be characterized as the function W for which the minimum of

$$\int_{\mathbb{R}^3 \diagdown \Omega_0} \left(|\nabla W|^2 + \frac{2}{r'} \frac{\partial W}{\partial \theta'} \frac{\partial V_0}{\partial r'} n \sin(n\theta') + 4 \cos(n\theta') \frac{\partial W}{\partial z'} \frac{\partial V_0}{\partial z'} \right) d\mathbf{x}'$$

is achieved or, equivalently, the solution to the boundary value problem

$$\begin{aligned} 2\Delta V_1 &= -\frac{2}{r'} \frac{\partial V_0}{\partial r'} n^2 \cos(n\theta') - 4 \frac{\partial^2 V_0}{\partial z'^2} \cos(n\theta') \quad \text{in } \mathbb{R}^3 \diagdown \Omega_0, \\ V_1 &= 0 \quad \text{at } \partial\Omega_0. \end{aligned}$$

We can find the solution to this problem in the form

$$V_1 = \cos(n\theta') \Phi(r', z'),$$

which leads to

$$C_{n,1} = \int_{-\infty}^{+\infty} \int_{a(z')}^{+\infty} \left(\frac{n^2}{2} \left| \frac{\partial V_0}{\partial r'} \right|^2 \right) r' dr' dz'$$

$$+ \int_{-\infty}^{+\infty} \int_{a(z')}^{+\infty} \left(\frac{1}{2} \left| \frac{\partial \Phi}{\partial r'} \right|^2 + \frac{1}{2} \left| \frac{\partial \Phi}{\partial z'} \right|^2 + \frac{n^2}{2r'^2} \Phi^2 - \frac{n^2}{r'} \Phi \frac{\partial V_0}{\partial r'} + 2 \frac{\partial \Phi}{\partial z'} \frac{\partial V_0}{\partial z'} \right) r' dr' dz'.$$

Hence, using $\Delta V_0 = 0$, one finds

$$-\Delta_{(r',z')} \Phi + \frac{n^2}{r'^2} \Phi = \frac{n^2}{r'} \frac{\partial V_0}{\partial r'} - \frac{2}{r'} \frac{\partial}{\partial r'} \left(r' \frac{\partial V_0}{\partial r'} \right) \quad \text{in } \mathbb{R}^3 \diagdown \Omega_0, \tag{4.1}$$

$$\Phi = 0 \quad \text{at } \partial\Omega_0. \tag{4.2}$$

We are going to show two important facts: (1) $C_{n,1}$ is strictly positive, and (2) $C_{n,1}$ is increasing with n. The first fact follows from the following calculations:

$$\begin{aligned} C_{n,1} &= \int_{-\infty}^{+\infty} \int_{a(z')}^{+\infty} \left(\frac{n^2}{2} \left| \frac{\partial V_0}{\partial r'} \right|^2 \right) r' dr' dz' + \int_{-\infty}^{+\infty} \int_{a(z')}^{+\infty} \left(\frac{1}{2} \left| \frac{\partial \Phi}{\partial z'} \right|^2 \right) r' dr' dz' \\ &\quad + \int_{-\infty}^{+\infty} \int_{a(z')}^{+\infty} \left(\frac{1}{2} \left| \frac{\partial \Phi}{\partial r'} \right|^2 + \frac{n^2}{2r'^2} \Phi^2 - \frac{n^2}{r'} \Phi \frac{\partial V_0}{\partial r'} + 2 \frac{\partial \Phi}{\partial z'} \frac{\partial V_0}{\partial z'} \right) r' dr' dz' \\ &> \int_0^H \int_{a(z')}^{+\infty} \left\{ \frac{n^2}{2} \left| \frac{\partial V_0}{\partial r'} \right|^2 + \frac{1}{2} \left| \frac{\partial \Phi}{\partial r'} \right|^2 + \frac{n^2}{2r'^2} \Phi^2 - \frac{n^2}{r'} \Phi \frac{\partial V_0}{\partial r'} - 2 \frac{\partial \Phi}{\partial r'} \frac{\partial V_0}{\partial r'} \right\} r' dr' dz' \\ &\underset{u=\log r}{=} \int_0^H \left[\int_{\log a(z')}^{+\infty} \left\{ \frac{n^2}{2} \left| \frac{\partial V_0}{\partial u} \right|^2 + \frac{1}{2} \left| \frac{\partial \Phi}{\partial u} \right|^2 + \frac{n^2}{2} \Phi^2 - n^2 \Phi \frac{\partial V_0}{\partial u} - 2 \frac{\partial \Phi}{\partial u} \frac{\partial V_0}{\partial u} \right\} du \right] dz'. \end{aligned}$$

By performing Fourier transform in u of this last expression we get

$$\begin{aligned}
&\int_0^H \left[\int_{-\infty}^{+\infty} \left\{\frac{n^2}{2}k^2 \left|\widehat{V_0}\right|^2 + \frac{1}{2}k^2 \left|\widehat{\Phi}\right|^2 + \frac{n^2}{2}\left|\widehat{\Phi}\right|^2 - ikn^2\overline{\widehat{\Phi}}\widehat{V_0} + 2k^2\overline{\widehat{\Phi}}\widehat{V_0}\right\} dk\right] dz' \\
&= \int_0^H \left\{\int_{-\infty}^{+\infty} \left[\frac{n^2}{2}k^2 \left|\widehat{V_0}\right|^2 + \left(\frac{1}{2}k^2 + n^2\right)\left|\widehat{\Phi}\right|^2\right.\right. \\
&\qquad \left.\left. + n^2k^2(\Im\widehat{V_0}\Re\widehat{\Phi} - \Im\widehat{\Phi}\Re\widehat{V_0}) + 2k^2((\Re\widehat{V_0}\Re\widehat{\Phi} + \Im\widehat{V_0}\Im\widehat{\Phi}))\right] dk\right\} dz. \qquad (4.3)
\end{aligned}$$

Let

$$\mathbf{a} = (\Re\widehat{\Phi}, \Im\widehat{\Phi}, \Re\widehat{V_0}, \Im\widehat{V_0}).$$

Then the integrand in (4.3) is $\mathbf{a}T(n,k)\mathbf{a}^t$ with

$$T(n,k) = \begin{pmatrix} \frac{k^2+n^2}{2} & 0 & k^2 & \frac{n^2k}{2} \\ 0 & \frac{k^2+n^2}{2} & -\frac{n^2k}{2} & k^2 \\ k^2 & -\frac{n^2k}{2} & \frac{k^2n^2}{2} & 0 \\ \frac{n^2k}{2} & k^2 & 0 & \frac{k^2n^2}{2} \end{pmatrix}$$

possessing the following two double eigenvalues:

$$\lambda_\pm = \frac{1}{4}k^2n^2 + \frac{1}{4}k^2 + \frac{1}{4}n^2 \pm \frac{1}{4}\sqrt{k^4n^4 - 2k^4n^2 + 17k^4 + 2k^2n^4 + 2k^2n^2 + n^4}.$$

Notice that

$$\lambda_- = \frac{k^4\left(n^2-4\right)}{k^2n^2 + k^2 + n^2 + \sqrt{k^4n^4 - 2k^4n^2 + 17k^4 + 2k^2n^4 + 2k^2n^2 + n^4}} > \frac{k^4/2}{1+k^2}\frac{n^2-4}{n^2+1}$$

so that

$$C_{n,1} > \frac{n^2-4}{n^2+1}\int_0^H \left[\int_{-\infty}^{+\infty} \frac{1}{2}\frac{k^4}{1+k^2}\left|\widehat{V_0}\right|^2 dk\right] dz = c'\frac{n^2-4}{n^2+1}.$$

The capacity is increasing with n:

$$\begin{aligned}
C_{n,1} &= \int_{-\infty}^{+\infty}\int_{a(z')}^{+\infty} \left(\frac{1}{2}\left|\frac{\partial\Phi}{\partial r'}\right|^2 + \frac{1}{2}\left|\frac{\partial\Phi}{\partial z'}\right|^2 + \frac{n^2}{2}\left|\frac{\Phi}{r'} - \frac{\partial V_0}{\partial r'}\right|^2 + 2\frac{\partial\Phi}{\partial z'}\frac{\partial V_0}{\partial z'}\right) r'dr'dz' \\
&\geq \int_{-\infty}^{+\infty}\int_{a(z')}^{+\infty} \left(\frac{1}{2}\left|\frac{\partial\Phi}{\partial r'}\right|^2 + \frac{1}{2}\left|\frac{\partial\Phi}{\partial z'}\right|^2 + \frac{(n-1)^2}{2}\left|\frac{\Phi}{r'} - \frac{\partial V_0}{\partial r'}\right|^2 + 2\frac{\partial\Phi}{\partial z'}\frac{\partial V_0}{\partial z'}\right) r'dr'dz' \\
&\geq \frac{1}{2}\min_{\Psi}\int_{-\infty}^{+\infty}\int_{a(z')}^{+\infty} \left(\left|\frac{\partial\Psi}{\partial r'}\right|^2 + \left|\frac{\partial\Psi}{\partial z'}\right|^2 + (n-1)^2\left|\frac{\Psi}{r'} - \frac{\partial V_0}{\partial r'}\right|^2 + 4\frac{\partial\Psi}{\partial z'}\frac{\partial V_0}{\partial z'}\right) r'dr'dz' \\
&= C_{n-1,1},
\end{aligned}$$

with equality if and only if

$$\Phi = r'\frac{\partial V_0}{\partial r'}, \qquad (4.4)$$

which would imply, by (4.1),

$$-\Delta_{(r',z')}\Phi + \frac{2}{r'}\frac{\partial \Phi}{\partial r'} = 0. \tag{4.5}$$

Multiplying (4.5) by Φ and integrating by parts using $\Phi = 0$ at $\partial\Omega$, we get

$$\int |\nabla \Phi|^2 \, d\mathbf{x}' = 0,$$

which would imply $\Phi = 0$ at almost every point, a fact that is incompatible with (4.4). Therefore the capacity is strictly increasing with n.

The fact that the capacity is strictly increasing with n is crucial to the proof of existence of symmetry-breaking bifurcations in section 3.3. Another important consequence of the result proved in this section, discussed in section 3.3, is the fact that perturbations with high order modes (large value of n) may be the most favorable energetically if the body of revolution that we perturb is sufficiently flat. This may lead to instabilities of the contact line in the form of numerous fingers.

Acknowledgments. The authors thankfully acknowledge the computer resources provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa) and the Spanish Supercomputing Network.

REFERENCES

[1] A. A. Ashour, *On a transformation of coordinates by inversion and its application to electromagnetic induction in a thin perfectly conducting hemispherical shell*, Proc. London Math. Soc., 15 (1965), pp. 557–576.

[2] S. I. Betelú and M. A. Fontelos, *Spreading of a charged microdroplet*, Phys. D, 209 (2005), pp. 28–35.

[3] S. I. Betelú, M. A. Fontelos, U. Kindelán, and O. Vantzos, *Singularities on charged viscous droplets*, Phys. Fluids, 18 (2006), article 051706.

[4] S. I. Betelú, M. A. Fontelos, and U. Kindelán, *The shape of charged drops: Symmetry-breaking bifurcations and numerical results*, in Elliptic and Parabolic Problems, Progr. Nonlinear Differential Equations Appl. 63, Birkhäuser, Basel, 2005, pp. 51–58.

[5] D. Duft, T. Achtzehn, R. Müller, B. A. Huber, and T. Leisner, *Rayleigh jets from levitated microdroplets*, Nature, 421 (2003), p. 128.

[6] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[7] R. Finn, *Equilibrium Capillary Surfaces*, Springer-Verlag, New York, 1986.

[8] M. A. Fontelos and A. Friedman, *Symmetry-breaking bifurcations of charged drops*, Arch. Ration. Mech. Anal., 172 (2004), pp. 267–294.

[9] O. D. Kellogg, *Foundations of Potential Theory*, Dover, New York, 1969.

[10] N. S. Landkof, *Foundations of Modern Potential Theory*, Springer-Verlag, New York, 1972.

[11] G. Lippmann, *Relations entre les phénomènes électriques et capillaires*, Ann. Chim. Phys., 494 (1875).

[12] F. Mugele and J. C. Baret, *Electrowetting: From basics to applications*, J. Phys. Condens. Matter, 17 (2005), pp. R705–R774.

[13] F. Mugele and J. Buehrle, *Equilibrium drop surface profiles in electric fields*, J. Phys. Condens. Matter, 19 (2007), article 375112.

[14] G. Pólya and G. Szegö, *Inequalities for the capacity of a condenser*, Amer. J. Math., 67 (1945), pp. 1–32.

[15] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.

[16] C. Pozrikidis, *Boundary Integral Methods for Linearized Viscous Flow*, Cambridge Texts Appl. Math., Cambridge University Press, UK, 1992.

[17] Lord Rayleigh, *On the equilibrium of liquid conducting masses charged with electricity*, Phil. Mag., 14 (1882), pp. 184–186.

[18] H. A. Stone, A. D. Stroock, and A. Ajdari, *Engineering flows in small devices: Microfluidics toward a lab-on-a-chip*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 36, Palo Alto, CA, 2004, pp. 381–411.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 149–168

MATHEMATICS AND MONUMENT CONSERVATION: FREE BOUNDARY MODELS OF MARBLE SULFATION[*]

FABRIZIO CLARELLI[†], ANTONIO FASANO[‡], AND ROBERTO NATALINI[†]

Abstract. We introduce some free boundary problems which describe the evolution of calcium carbonate stones under the attack of atmospheric SO_2, taking into account both swelling of the external gypsum layer and the influence of humidity. Different behaviors are described according to the relative humidity of the environment, and in all cases reliable explicit quasi-steady approximations are introduced under reasonable assumptions on the data. Some numerical simulations are also performed to describe gypsum formation using experimental data, which show a good agreement with the quasi-steady solutions. The influence of the cleaning the crust and of the change in concentration of pollution is evaluated and discussed.

Key words. free boundary problems, chemical damage, porous media, swelling, influence of humidity, damage of cultural heritage

AMS subject classifications. Primary, 76S05; Secondary, 35R35

DOI. 10.1137/070695125

1. Introduction.

Deterioration of stones is a complex problem and one of the main concerns for people working in the field of conservation and restoration of cultural heritage. It is extremely difficult to isolate a single factor in these kinds of processes, which are the results of the interaction of various mechanisms, many of which also occur in natural weathering; however, atmospheric pollution can certainly be considered as one of the most important factor of damage. In this paper we shall introduce some free boundary models to describe damage induced by pollution.

Although in recent years air pollution in European urban areas has decreased considerably, there still remain concentrations of pollutants such as sulfur dioxide (SO_2) from combustion of fossil fuels, and nitrogen oxides (NO_x) from combustion engines, the former being the most important factor in the deterioration of stones. Indeed SO_2 can react with any calcareous component in the stone, producing an external layer of gypsum, which may be drained away by rain or form crusts that eventually exfoliate. This process greatly depends on the nature of the stone and on the presence of moisture. Since the stone is a porous material, condensation of moisture may occur deep within the pores of the material, and it is critical to its reactivity to pollutant. Despite the intense experimental research performed in this area (see [5] for a large review), further studies are necessary in order to provide a predictive tool.

The present paper will be concerned with the so-called dry deposition of SO_2 on calcium carbonate stones. Many experimental investigations (see, for instance, [7, 5]) have shown that this process, which is mainly influenced by short-range transport of pollutants from local sources, is the main source of damage in stones, and more precisely in very compact stones like high-quality marbles. Wet deposition, in which

[*] Received by the editors October 17, 2007; accepted for publication (in revised form) May 19, 2008; published electronically September 8, 2008.

http://www.siam.org/journals/siap/69-1/69512.html

[†] Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, I–00161 Rome, Italy (clarellif@gmail.com, r.natalini@iac.cnr.it).

[‡] Dipartimento di Matematica "Ulisse Dini," Università degli Studi di Firenze, Viale Morgagni 67/A, I-50134 Florence, Italy (fasano@math.unifi.it).

pollutants are dissolved in moisture droplets and rain, is actually considered as secondary, even if, as in areas where buildings and monuments remain wet for a long time, it may become important.

For dry deposition the path of reaction of SO_2 with calcium carbonate is revealed by the x-ray diffraction counts, which suggest that the reaction can be approximated by the following simplified one-step reaction [2, 7]:

$$CaCO_3 + SO_2 + \frac{1}{2}O_2 + 2H_2O \to CaSO_4 \cdot 2H_2O + CO_2. \tag{1.1}$$

Namely, one mole of calcium carbonate and one mole of sulfur dioxide, combined with two moles of water, produce one mole of calcium sulfate dihydrate (gypsum) $CaSO_4 \cdot 2H_2O$ and one of carbon dioxide CO_2. In the following we will use the fact that, on the typical time scale of the whole process, which will be one year, not only can we neglect the intermediate steps leading to (1.1), but we may also consider this reaction instantaneous, thus producing a sharp free boundary between gypsum and the unreacted calcium carbonate. In stones with very low porosity, the appearance of a sharp gypsum-marble interface has been experimentally observed in [7, 4, 8]. Moreover, according to the results in [9], this is an excellent approximation of the finite rate regime.

In our model we will consider only a one-dimensional geometry. Actually, since the typical thickness of the gypsum layer produced in one year in standard conditions is 20 [μm], if the gypsum layer is not removed, then for surfaces without too high a curvature a one-dimensional model is fully appropriate. However, a peculiar feature of the process is that the advancement of the sulfation front depends on factors that can change several hundreds of times during the same time scale. Such factors may be influenced very much by local conditions, with the consequence that monuments made of the same material, having the same age, and located in the same city not far from each other can have a very different state of preservation. Thus, even if we consider a small number of influencing factors, the resulting picture is not simple and requires a rather detailed knowledge of data concerning not only the regional climatic conditions, but also the local environment.

An additional comment about reaction (1.1) is in order. As a matter of fact, in real processes, calcium sulfate is also produced, according to the reaction path

$$CaCO_3 + SO_2 + \frac{1}{2}H_2O \to CaSO_3 \cdot \frac{1}{2}H_2O + CO_2.$$

In this regard, it has been shown that the proportion of $CaSO_4$ and $CaSO_3$ in a sulfation process can be affected to a great extent not only by relative humidity, but also by the presence of other substances like $CaCl_2$, $MnCl_2$, $CuCl_2$, and $FeCl_3$ acting as catalysts which favor one of the two reactions [7, 11]. The two reactions generally occur simultaneously and the resulting mixed layer can have properties (e.g., porosity, permeability, etc.) depending on the volume fractions of the two substances, and calcium sulfite may be converted to calcium sulfate, so that the process, including the two reactions, is considerably more complicated. However, even if laboratory tests show that calcium sulfite is the primary product of the reaction, in situ analysis reveals only the presence of calcium sulfate, which can be considered as the final state reached by the whole reaction.

In this paper we present some free boundary models, which include important phenomena not considered in the previous mathematical papers on the subject [3, 1]:

swelling and relative humidity. As we can see in the following, these two factors have a deep influence in the evolution of sulfation and require specific consideration. Using dimensional scaling we are able to build almost explicit approximations of these models, which will be very useful in their calibration and qualitative study. Different regimes are determined, according to the presence of relative humidity near the stones. We also introduce some finite difference schemes to compare the results of our asymptotic analysis and the real behavior of the solutions. A good agreement is found in all cases, even if on the scale of 10 years the numerical schemes should be preferred. Finally, some simulations are performed utilizing real data, which were kindly provided by Arpalazio, the Rome regional authority for monitoring pollution. Useful indications are derived by our elaboration of these data.

2. Swelling. The transformation of marble into gypsum is accompanied by a volume change. The swelling rate can be calculated easily because the molar ratio in (1.1) between $CaCO_3$ and $CaSO_4$ is 1 : 1. Thus, on the unit surface of the reaction front the consumption rate of $CaCO_3$ moles equals the production rate of $CaSO_4$ moles. If $x = \sigma(t)$ denotes the sulfation front and $x = \sigma_0(t)$ is the gypsum surface exposed to air (the frame of reference is chosen so that unreacted marble is at rest), we have the relation

$$\dot{\sigma}_0 = -\omega\dot{\sigma}, \tag{2.1}$$

where $\omega = \frac{\mu_m}{\mu_s} - 1$, μ_m and μ_s being the molar density (# mol/cm^3) of $CaCO_3$ in the marble and of $CaSO_4$ in the gypsum, respectively. We are supposing that μ_m is constant (i.e., the marble is a homogeneous material) and that μ_s is also constant, meaning that gypsum is formed with some standard structure, independently of its production rate. Under this assumption, if $\sigma_0(0) = \sigma(0) = 0$, we may conclude that

$$\sigma_0(t) = -\omega\sigma(t), \tag{2.2}$$

so that the thickness of the gypsum layer at time t is

$$h(t) = (1 + \omega)\sigma(t). \tag{2.3}$$

Otherwise if $\sigma(0) > 0$ and $\sigma_0(0) > 0$, we obtain

$$\sigma_0(t) - \sigma_0(0) = -\omega\left(\sigma(t) - \sigma(0)\right). \tag{2.4}$$

Swelling is an important phenomenon. First of all *it is not small*. Although the determination of μ_s is not easy, it is reasonable to say that the swelling rate ω can approach 2. The motion of gypsum influences the flow of the air and of the other gaseous components present in the pores. We concentrate our attention on SO_2 and water vapor. On the length and time scales typical of the process, air can be considered to move with the same speed as gypsum, i.e.,

$$v_a = \dot{\sigma}_0. \tag{2.5}$$

This approximation will be completely justified in the appendix.

3. Sulfation: A two-regime process. Experimental researchers know that relative humidity has a key role in regulating the speed of marble sulfation. One could think that since SO_2 is by far the most diluted among the reactants in (1.1), it has to play a limiting role. On the contrary, this role is taken up by H_2O. Indeed it is observed that when relative humidity exceeds some threshold, then SO_2 reacts completely. Below that threshold (which is around 75%) there is another range (down to 45%) in which the reaction slows down and stops completely for even lower values; see [5, 7] and the references therein. We can interpret this phenomenon as follows.

According to (1.1) a molecule of SO_2 coming in contact with $CaCO_3$ reacts if two molecules of H_2O are available at the same point (we suppose that there is always enough O_2). Such a multiple encounter has a negligibly small probability to occur if H_2O is just in the gaseous form. To make the reaction proceed at full speed it is necessary that H_2O is permanently available. This is true only if vapor condenses on the unreacted marble surface forming a liquid film.

Condensation is made possible by the fact that marble is hygroscopic and will be the result of a sorption-desorption process, which we assume to go through equilibrium states. Therefore, we may interpret the limiting role of H_2O by saying that the liquid film is present if the relative humidity is above the 75% threshold, while in the range 45%–75% there will be just humid spots on which the reaction takes place. Accordingly only a fraction of the SO_2 arriving to the front will be employed in the reaction. For relative humidity below 45% no liquid water is present and the reaction stops.

The two sulfation regimes (full and reduced speed) have different boundary conditions on the unreacted marble surface. We will examine them separately.

4. Full sulfation speed. If temperature T and pressure p are prescribed, the condition for full reaction speed is that the concentration of H_2O in air exceeds some value $w_0(T,p)$.

Let s denote the concentration of SO_2 in the pores of gypsum. The flow of SO_2 relative to air is governed by Fick's law. Thus in the frame of reference where marble is at rest the SO_2 flux has the expression

$$j_s = n_g\Big(-d_s\frac{\partial s}{\partial x} - s\omega\dot{\sigma}\Big), \tag{4.1}$$

n_g denoting gypsum porosity and d_s the diffusivity of SO_2 in air.

We have used (2.5) and (2.1). Consistently with the assumption $\mu_s =$ constant, we suppose also $n_g =$ constant.

Thus the SO_2 mass balance in the gypsum layer $\sigma_0(t) < x < \sigma(t)$ is expressed by

$$\frac{\partial s}{\partial t} - d_s\frac{\partial^2 s}{\partial x^2} - \omega\dot{\sigma}\frac{\partial s}{\partial x} = 0. \tag{4.2}$$

The value of s at the external boundary is some known function of time

$$s(\sigma_0(t), t) = s_a(t). \tag{4.3}$$

When SO_2 reacts totally at the front we have

$$s(\sigma(t), t) = 0, \tag{4.4}$$

implying that the flux of SO_2 at the front is purely diffusive. Hence the mass balance in the reaction is

$$-n_g\frac{d_s}{M_s}\frac{\partial s}{\partial x}(\sigma(t), t) = \frac{\rho_m}{M_m}\dot{\sigma}, \tag{4.5}$$

where M_s, M_m are the molar weights of SO_2 and of $CaCO_3$, respectively, and ρ_m is the density of the pristine marble (the ratio $\frac{\rho_m}{M_m}$ is nothing but the molar density μ_m we have already introduced).

Thus in this region the problem for the pair (s, σ) can be formulated independently of the evolution of other quantities.

Although the water vapor concentration w plays no role during this time, it is important to monitor its evolution in view of the possible transition to the other regime, since the constraint

$$w(\sigma(t), t) \geq w_0(T, p) \tag{4.6}$$

must be satisfied during this stage.

The water vapor flux is

$$j_w = n_g \Big(-d_w \frac{\partial w}{\partial x} - w\omega\dot{\sigma} \Big) \tag{4.7}$$

(d_w is the diffusivity of H_2O in air), thus the H_2O mass balance is

$$\frac{\partial w}{\partial t} - d_w \frac{\partial^2 w}{\partial x^2} - \omega\dot{\sigma} \frac{\partial w}{\partial x} = 0. \tag{4.8}$$

At the outer surface w equals the external concentration (see the comment about (4.3)).

$$w(\sigma_0(t), t) = w_a(t). \tag{4.9}$$

On the free boundary

$$-n_g \frac{d_w}{M_w} \frac{\partial w}{\partial x} = 2 \frac{\rho_m}{M_m} \dot{\sigma} + n_g (1 + \omega) \frac{w}{M_w} \dot{\sigma} \tag{4.10}$$

(M_w molar weight of H_2O), since two moles of H_2O react with one mole of $CaCO_3$.

The validity of (4.6) must be checked at all times. The system (4.2)–(4.5) can be reduced to a standard Stefan problem, as we shall see, and once $\sigma(t)$ is known, problem (4.8)–(4.10) presents no difficulties.

Remark 4.1. The balance in (4.10) contains some implicit assumptions. As we have stated, in order to trigger the reaction, water must condense as a liquid film. The production of $CaSO_4$ is the result of intermediate reactions occurring in the film and producing H_2SO_4, which eventually reacts with $CaCO_3$. In writing (4.10) we neglected the film thickness and also the moisture content in the pristine marble. Strictly speaking, the left-hand side (l.h.s.) in (4.10) must be interpreted as the feeding rate of the water film, whose thickness we suppose to be very small in comparison with the typical scale length of the process and constant. The first term on the right-hand side (r.h.s.) of (4.10) is the water moles consumption rate in the reaction per unit surface. However, if the pores of the pristine marble contain some water, this would enter the balance with a term $n_m \dot{\sigma} \frac{w_m}{M_w}$ (n_m marble porosity, w_m water concentration in marble pores).

If by chance marble is saturated by liquid water ($w_m =$ liquid water density), this term may not be negligible. Moreover, the water within the marble may contribute to keep a sufficiently high relative humidity in the gypsum for some time, even when the relative humidity in air drops below the full speed threshold. In any case the fact that marble is hygroscopic is going to affect the water vapor flow within the marble. However, in the case of stones having a very low porosity, the mass fraction of the water possibly stored in the pores is not enough to maintain a liquid film at the reaction front when air humidity is below threshold. In any case, the threshold values (45% and 75%) are currently found in the literature (with some fluctuation), but the state of the stone to which they refer is never specified. So they may already include the effect of stored water.

5. Reduced sulfation speed. We have now w between two thresholds:

$$w_1(T,p) \le w(\sigma(t),t) \le w_0(T,p). \tag{5.1}$$

The governing differential equations for s, w remain unchanged (i.e., (4.2), (4.8)), as do the conditions on the external boundary (4.3), (4.9). A deep modification intervenes on the free boundary, since the $CaCO_3$ front is not coated by a continuous water film, but rather covered by humid spots. We can define an efficiency factor $\alpha(w,T,p)$ (which for simplicity we denote $\alpha(w) = \frac{1}{w_0-w} - \frac{1}{w_0-w_1}$) for the chemical reaction, which increases from 0 to ∞ as w goes from w_1 (the no-reaction threshold) to w_0 (the full reaction threshold). The two free boundary conditions (4.4), (4.5) are now replaced by

$$\frac{j_s}{M_s} = \frac{\rho_m}{M_m}\dot\sigma + n_g \frac{s}{M_s}\dot\sigma, \tag{5.2}$$

$$\frac{\rho_m}{M_m}\dot\sigma = n_g \alpha(w) \frac{s}{M_s}. \tag{5.3}$$

The first equation expresses the total molar balance of SO_2, including the loss rate due to the reaction and the advective flux due to the transport of the residual SO_2 by the moving front. The second condition contains the factor $\alpha(w)$ specifying the reaction efficiency (α is dimensionally a velocity). When α goes to $+\infty$, $s(\sigma(t),t)$ is forced to tend to zero and we are back to the full speed regime. When α vanishes the front stops and the SO_2 flux vanishes too, yielding $\frac{\partial s}{\partial x} = 0$.

6. Lagrangian coordinate. Before we deal with the flow of air, it is convenient to adopt a frame of reference (ξ,t) moving with the gypsum. In the frame (x,t) we have used so far the marble is at rest. Let us consider a gypsum particle which is formed at the point $x=\xi$ at a time $\tau(\xi)$. Following its motion up to the time t, we find it at the location

$$x = \xi + \int_{\tau(\xi)}^{t} \dot\sigma_0(\vartheta)d\vartheta = \xi - \omega[\sigma(t) - \sigma(\tau(\xi))] = (1+\omega)\xi - \omega\sigma(t), \tag{6.1}$$

since by definition $\sigma(\tau(\xi)) = \xi$. Thus ξ plays the role of a Lagrangian coordinate. Inverting (6.1) we find

$$\xi = \frac{1}{1+\omega}x + \frac{\omega}{1+\omega}\sigma(t) \tag{6.2}$$

and of course $x = \xi = \sigma(t)$ on the free boundary. We define

$$s(\xi,t) = s(x,t), \tag{6.3}$$

so that the domain $\sigma_0(t) < x < \sigma(t)$, $t > 0$, is mapped to $0 < \xi < \sigma(t)$, $t > 0$, and (4.2)–(4.5) transform to

$$\frac{\partial s}{\partial t} - \frac{d_s}{(1+\omega)^2}\frac{\partial^2 s}{\partial \xi^2} = 0, \tag{6.4}$$

$$s(0,t) = s_a(t), \tag{6.5}$$

$$s(\sigma(t),t) = 0, \tag{6.6}$$

$$-n_g \frac{d_s}{M_s}\frac{1}{1+\omega}\frac{\partial s}{\partial \xi} = \frac{\rho_m}{M_m}\dot{\sigma}. \tag{6.7}$$

Similarly, we introduce

$$w(\xi,t) = w(x,t) \tag{6.8}$$

and (4.8)–(4.10) become

$$\frac{\partial w}{\partial t} - \frac{d_w}{(1+\omega)^2}\frac{\partial^2 w}{\partial \xi^2} = 0, \tag{6.9}$$

$$w(0,t) = w_a(t), \tag{6.10}$$

$$-n_g \frac{d_w}{M_w}\frac{1}{1+\omega}\frac{\partial w}{\partial \xi} = 2\frac{\rho_m}{M_m}\dot{\sigma} + n_g(1+\omega)\frac{w}{M_w}\dot{\sigma}. \tag{6.11}$$

The free boundary conditions (5.2), (5.3) for the reduced speed regime take the form

$$-\frac{1}{M_s}\Big(\frac{d_s}{1+\omega}\frac{\partial s}{\partial \xi} + n_g\omega\dot{\sigma}s\Big) = \Big(\frac{\rho_m}{M_m} + n_g\frac{s}{M_s}\Big)\dot{\sigma}, \tag{6.12}$$

$$\frac{\rho_m}{M_m}\dot{\sigma} = n_g\alpha(w)\frac{s}{M_s}. \tag{6.13}$$

7. Rescaling. Let σ^*, t^* be suitable length and time scales. Set $\eta = \xi/\sigma^*$, $\vartheta = t/t^*$, $\delta(\vartheta) = \sigma(t^*\vartheta)/\sigma^*$, and define

$$\widehat{s}(\eta,\vartheta) = s(\xi,t)/s^*, \qquad \widehat{w}(\eta,\vartheta) = w(\xi,t)/w^*.$$

To be specific, we take $t^* = 1$ year $\simeq 3.15 \cdot 10^7$ sec and $\sigma^* = 2 \cdot 10^{-3}$ cm. Then we take $s^* = 14.3 \cdot 10^{-12}\mathrm{g}\cdot\mathrm{cm}^{-3}$ as a typical yearly average of SO_2 concentration in air, and $w^* = 13 \cdot 10^{-6}\mathrm{g}\cdot\mathrm{cm}^{-3}$ as the H_2O concentration coinciding with the threshold w_0 corresponding to P_0 and to a fixed temperature $T = 20°$ Celsius.

7.1. Rescaling the SO_2 flow problem. The system (6.4)–(6.7) rescales to

$$\frac{\partial \widehat{s}}{\partial \vartheta} - \frac{1}{(1+\omega)^2}K_s\frac{\partial^2 \widehat{s}}{\partial \eta^2} = 0, \tag{7.1}$$

$$\widehat{s}(0,\vartheta) = \widehat{s}_a(\vartheta), \tag{7.2}$$

$$\widehat{s}(\delta(\vartheta),\vartheta) = 0, \tag{7.3}$$

$$-n_g K_s\frac{1}{1+\omega}\frac{s^*}{M_s}\frac{M_m}{\rho_m}\frac{\partial \widehat{s}}{\partial \eta} = \frac{d\delta}{d\vartheta}, \tag{7.4}$$

where $\widehat{s}_a(\vartheta) = s_a(t^*\sigma)/s^*$, $K_s = \frac{t^* d_s}{\sigma^{*2}}$, and $d_s = 0.1$ cm^{2}sec^{-1}. With $\omega \simeq 2$ we have $\frac{K_s}{(1+\omega)^2} \simeq 9 \cdot 10^{10} \gg 1$ and here too (7.1) simplifies to $\frac{\partial^2 \widehat{s}}{\partial \eta^2} \simeq 0$. With $n_g = 0.3$, the coefficient of $\frac{\partial \widehat{s}}{\partial \eta}$ in (7.4) is

$$\Omega_s = \frac{n_g}{1+\omega} \frac{s^* M_m}{M_s \rho_m} K_s \simeq 0.286. \tag{7.5}$$

Thus we may say that at each time $\widehat{s}(\eta, \vartheta)$ is very well approximated by a linear function of η,

$$\widehat{s}(\eta, \vartheta) = \widehat{s}_a(\vartheta) - \gamma(\vartheta)\eta, \tag{7.6}$$

satisfying the conditions

$$\gamma(\vartheta)\delta(\vartheta) = \widehat{s}_a(\vartheta), \qquad \Omega_s \gamma(\vartheta) = \dot{\delta}(\vartheta).$$

Hence we obtain

$$\delta(\vartheta) = \Big[2\Omega_s \int_0^\vartheta \widehat{s}_a(\tau) d\tau\Big]^{1/2}, \qquad \widehat{s}(\eta, \vartheta) = \widehat{s}_a(\vartheta)\Big[1 - \frac{\eta}{\delta(\vartheta)}\Big]. \tag{7.7}$$

The procedure of approximating $\widehat{s}$ as in (7.6) is justified as long as $\frac{\partial \widehat{s}}{\partial \vartheta}$ is not singular, or more precisely as long as

$$\frac{\partial \widehat{s}}{\partial \vartheta} \cdot 10^{-11} \ll 1 \tag{7.8}$$

in our setting. However, if $\widehat{s}_a(0)$ is not zero, we know that (7.1)–(7.4) has an explicit solution with $\frac{d\delta}{d\vartheta} \approx \frac{1}{\sqrt{\vartheta}}$ and $\frac{\partial \widehat{s}}{\partial \vartheta} \approx \frac{1}{\vartheta}$, which could make approximation not applicable. Let us investigate these points more carefully. Let us consider the case $\widehat{s}_a(\vartheta) = \widehat{s}_0 > 0$. Set $A = \frac{K_s}{(1+\omega)^2}$. The explicit solution of (7.1)–(7.4) is

$$\widehat{s}(\eta, \vartheta) = \frac{A}{\Omega_s} 2\gamma e^{\gamma^2} \int_{\frac{\eta}{2\sqrt{A\vartheta}}}^{\gamma} e^{-\xi^2} d\xi, \qquad \delta(\vartheta) = 2\gamma\sqrt{A\vartheta}, \tag{7.9}$$

where γ is the unique solution of

$$\frac{\Omega_s}{A} \widehat{s}_0 = 2\gamma e^{\gamma^2} \int_0^\gamma e^{-\xi^2} d\xi. \tag{7.10}$$

Now, we notice that $\Omega_s / A \approx 3 \cdot 10^{-11}$, implying that $\gamma \ll 1$. Therefore, setting $F(\gamma) = 2\gamma e^{\gamma^2} \int_0^\gamma e^{-\xi^2} d\xi$, since $F(0) = \partial_\gamma F(0) = 0$ and $\partial_{\gamma\gamma} F(0) = 4$, we may approximate the r.h.s. of (7.10) as $F(\gamma) \simeq 2\gamma^2$, and hence

$$\gamma \simeq \left(\widehat{s}_0 \frac{\Omega_s}{2A}\right)^{1/2} \quad \left[\simeq 4 \cdot 10^{-6}\right], \tag{7.11}$$

concluding that

$$\delta(\vartheta) \simeq \sqrt{2\Omega_s \widehat{s}_0 \vartheta}. \tag{7.12}$$

This is precisely formula (7.7), which is therefore justified even for small ϑ. At the same time we conclude that

$$\widehat{s}(\eta, \vartheta) = \widehat{s}_0 \left(1 - \frac{\eta}{\sqrt{2\Omega_s \vartheta \widehat{s}_0}}\right), \quad 0 < \eta < \sqrt{2\Omega_s \vartheta \widehat{s}_0}. \tag{7.13}$$

7.2. Rescaling the $\mathbf{H_2O}$ flow problem. The system (6.9)–(6.11) rescales to

$$\frac{\partial \widehat{w}}{\partial \vartheta} - \frac{K_w}{(1+\omega)^2}\frac{\partial^2 \widehat{w}}{\partial \eta^2} = 0, \qquad K_w = \frac{t^* d_w}{\sigma^{*2}} = K_s \frac{d_w}{d_s} \gg 1, \tag{7.14}$$

$$\widehat{w}(0,\vartheta) = \widehat{w}_a(\vartheta) = w_a(t^*\vartheta)/w^*, \tag{7.15}$$

$$-\Omega_w \frac{\partial \widehat{w}}{\partial \eta} = \Big[1 + \frac{1}{2} n_g (1+\omega) \frac{w^* M_m \widehat{w}}{M_w \rho_m}\Big] \frac{d\delta}{d\vartheta}, \tag{7.16}$$

with

$$\begin{aligned}\Omega_w &= \frac{1}{2}\frac{n_g}{1+\omega}\frac{d_w w^* t^*}{M_w \sigma^{*2}}\frac{M_m}{\rho_m}\\ &= \frac{1}{2}\frac{n_g}{1+\omega}\frac{w^* M_m}{M_w \rho_m} K_w = \frac{1}{2}\frac{w^*}{s^*}\frac{M_s}{M_w}\frac{d_w}{d_s}\Omega_s.\end{aligned}$$

Due to the smallness of the ratio $\frac{w^*}{\rho_m}$, condition (7.16) can be simplified to

$$-\Omega_w \frac{\partial \widehat{w}}{\partial \eta} = \frac{d\delta}{d\vartheta}. \tag{7.17}$$

Then, according to the approximation $\widehat{w}_{\eta\eta} = 0$,

$$\widehat{w}(\eta,\vartheta) = \widehat{w}_a(\vartheta) - \beta(\vartheta)\eta, \tag{7.18}$$

where $\beta = -\widehat{w}_\eta = \dot{\delta}/\Omega_w = \frac{\Omega_s}{\Omega_w}\frac{\widehat{s}_a(\vartheta)}{\delta(\vartheta)}$. Then

$$\widehat{w}(\eta,\vartheta) = \widehat{w}_a(\vartheta) - \frac{\Omega_s}{\Omega_w}\frac{\widehat{s}_a(\vartheta)}{\delta(\vartheta)}\eta. \tag{7.19}$$

All the coefficients used in (7.1), (7.4), (7.14), (7.16) are displayed in Table 1. In addition, Table 2 shows the reference values of the parameters used in our considerations as well as in the numerical simulations in the following sections.

It is useful to observe that the different order of magnitude of $\Omega_s = 0.286$ and $\Omega_w \simeq 3 \cdot 10^6$, related to the ratio $\frac{w^*}{s^*}$, yields a strong qualitative difference between the transport of SO_2 and H_2O within gypsum.

TABLE 1
Rescaling.

$K_s = \frac{t^* d_s}{\sigma^{*2}}$
$K_w = K_s \frac{d_w}{d_s}$
$\Omega_s = \frac{n_g}{1+\omega}\frac{s^* M_m}{M_s \rho_m} K_s$
$\Omega_w = \frac{1}{2}\frac{n_g}{1+\omega}\frac{d_w M_m w^* t^*}{M_w \rho_m \sigma^{*2}}$

TABLE 2
Values of parameters.

Parameter	Meaning	Dimensions	value
n_g	Gypsum porosity	nondimensional	0.3
n_m	Marble porosity	nondimensional	0.005–0.015
M_w	Molar weight (water)	[g/mol]	18.0153
M_m	Molar weight (marble)	[g/mol]	100.087
M_m	Molar weight (gypsum)	[g/mol]	172.166
d_s	Diffusivity in gypsum (SO_2)	[cm^2/sec]	0.1
d_w	Diffusivity in gypsum (H_2O)	[cm^2/sec]	0.2178
ρ_m	mass density of marble	[g/cm^3]	2.83
ρ_w	mass density of water	[g/cm^3]	1
ρ_g	mass density of gypsum	[g/cm^3]	1.6
ω	molar density ratio	nondimensional	≈ 2
σ^*	reference layer in 1 year	[cm]	$2 \cdot 10^{-3}$
t^*	reference time (1 year)	[sec]	$3.15 \cdot 10^{7}$
s^*	reference density (SO_2)	[g/cm^3]	$14.3 \cdot 10^{-12}$
w^*	reference density (H_2O)	[g/cm^3]	$17.3 \cdot 10^{-6}$

7.3. Rescaling the SO_2 and H_2O flow in the reduced speed regime. The rescaled version of (6.12), (6.13) is

$$-\Omega_s \frac{\partial \widehat{s}}{\partial \eta} = \frac{d\delta}{d\vartheta}\Big[1 + n_g \frac{s^* M_m \widehat{s}}{\rho_m M_s}(1+\omega)\Big], \tag{7.20}$$

$$\frac{d\delta}{d\vartheta} = n_g \frac{s^* M_m}{M_s \rho_m}\widehat{\alpha}(\widehat{w})\widehat{s} = \lambda_1 \widehat{\alpha}(\widehat{w})\widehat{s}, \qquad \widehat{\alpha} = \frac{\alpha}{v^*}. \tag{7.21}$$

Again we may simplify (7.20) to

$$-\Omega_s \frac{\partial \widehat{s}}{\partial \eta} = \frac{d\delta}{d\vartheta}. \tag{7.22}$$

The solution (7.6) on the boundary $\eta = \delta(\vartheta)$ is

$$\widehat{s}(\delta, \vartheta) = \widehat{s}_a(\vartheta) - \frac{d\delta}{d\vartheta}\frac{\delta(\vartheta)}{\Omega_s}; \tag{7.23}$$

substituting (7.21) in (7.23), we get

$$\widehat{s}(\delta, \vartheta) = \frac{\widehat{s}_a(\vartheta)}{\left(1 + \dfrac{\lambda_1}{\Omega_s}\widehat{\alpha}\delta(\vartheta)\right)}. \tag{7.24}$$

Then, using (7.24) in (7.21), we have

$$\frac{d\delta}{d\vartheta} = \lambda_1 \widehat{\alpha} \frac{\widehat{s}_a(\vartheta)}{\left(1 + \dfrac{\lambda_1}{\Omega_s}\widehat{\alpha}\delta(\vartheta)\right)}; \tag{7.25}$$

finally, from (7.6) and (7.18)

$$\widehat{s}(\eta,\vartheta)=\widehat{s}_a(\vartheta)-\frac{d\delta}{d\vartheta}\frac{\eta}{\Omega_s}, \tag{7.26}$$

$$\widehat{w}(\eta,\vartheta)=\widehat{w}_a(\vartheta)-\frac{d\delta}{d\vartheta}\frac{\eta}{\Omega_w}. \tag{7.27}$$

The initial condition depends on when the reduced speed regime is started. Note that when $\widehat{\alpha}$ becomes large, (7.25) reduces to $\dot{\delta}\delta=\Omega_s\widehat{s}_a(\vartheta)$, i.e., the full speed law of advancement.

Remark 7.1. A possible variation in time of the water film thickness would imply an additional term. This is not a trivial change. Suppose, for instance, that the water film thickness is $\varepsilon(w)$ with $\varepsilon=0$ below the lower threshold, $\varepsilon=\varepsilon_{\max}$ above the upper threshold, and $\frac{d\varepsilon}{dt}>0$ in between. In the range of w where ε varies the additional water molar flux created by the variation of the water film thickness is $n_g\frac{d\varepsilon}{dt}\frac{\rho_w}{M_w}$, where ρ_w is the density of liquid water. This means that the term to be added on the r.h.s. of (4.10) is

$$n_g\frac{\rho_w}{M_w}\frac{d\varepsilon}{dt}\Big(\frac{\partial w}{\partial t}+\dot{\sigma}\frac{\partial w}{\partial x}\Big),$$

which makes the problem of finding w much more difficult. However, we point out that performing the rescaling so far adopted, the influence of this effect can be neglected.

Remark 7.2. We remark that in (7.1) and (7.14), the coefficients K_s and K_w play the same role as the nondimensional number $(Ste)^{-1}$ in phase change problems. With *Ste* we indicate the Stefan number; see, for instance, [6]. Our approximation corresponds to the case $Ste\ll 1$ (latent heat dominant with respect to the so-called sensible heat) and makes the quasi-steady approach feasible. We refer to [6] for some results for small and large Stefan numbers.

8. Numerical schemes. As we have seen, the quasi-steady approximations provide a simple way of constructing solutions; however, it is advisable to set up a numerical scheme capable of dealing with the complete equations (i.e., including the inertia terms). The reason is twofold:

1. Although we have limited our attention to a one-year period (which is the time for which data were available to us), in practical cases one could be interested in predictions extended over a period of ten years or more. We will see that for the cases here examined the discrepancy between the full model and the quasi-steady approximation ranges between 2%–4%, but errors may accumulate over the years. Also, if we used data with higher frequency and amplitude, we could obtain a greater relative error.
2. It may happen that the SO_2 concentration and/or relative humidity undergo sudden changes (for instance, due to a massive air replacement in the region where the monument is situated). This event may not be correctly described on the basis of the quasi-steady approximation.

These arguments motivate the numerical scheme we are going to present in this section. The quasi-steady approximation will be useful anyway in providing a reliable starting point.

Consider equations (7.1), (7.2), (7.3), (7.4), (7.14), (7.15), (7.17), and (7.21). This is a problem with a moving boundary given by $\delta(\vartheta)$. We prefer to change coordinates, and so obtain a fixed boundary. To this purpose, we introduce (y,τ)

such that $y = \eta/\delta(\vartheta)$ and $\tau = \vartheta$, the domain changes to $0 \le y \le 1$, supposing $\delta(0) > 0$.

Equations change to

$$\frac{\partial \widehat{s}}{\partial \tau} - \frac{y\dot{\delta}(\tau)}{\delta(\tau)}\frac{\partial \widehat{s}}{\partial y} - \frac{1}{(1+\omega)^2(\delta(\tau))^2}K_s\frac{\partial^2 \widehat{s}}{\partial y^2} = 0, \tag{8.1}$$

$$\widehat{s}(y=0,\tau) = \widehat{s}_a(\tau), \tag{8.2}$$

$$-\Omega_s\frac{1}{\delta(\tau)}\frac{\partial \widehat{s}}{\partial y} = \frac{d\delta}{d\tau} \qquad \text{at} \quad y=1, \tag{8.3}$$

$$\widehat{s}(y=1,\tau) = 0 \quad \text{(full speed)}, \tag{8.4}$$

or

$$\frac{d\delta}{d\tau} = n_g\frac{s^* M_m}{M_s\rho_m}\widehat{\alpha}(\widehat{w})\widehat{s}, \qquad \widehat{\alpha} = \frac{\alpha}{v^*} \quad \text{(reduced speed)}, \tag{8.5}$$

$$\frac{\partial \widehat{w}}{\partial \tau} - \frac{y\dot{\delta}(\tau)}{(\delta(\tau))^2}\frac{\partial \widehat{w}}{\partial y} - \frac{K_w}{(1+\omega)^2\delta(\tau)^2}\frac{\partial^2 \widehat{w}}{\partial y^2} = 0, \tag{8.6}$$

$$\widehat{w}(y=0,\tau) = \widehat{w}_a(\tau), \tag{8.7}$$

$$-\frac{\Omega_w}{\delta(\tau)}\frac{\partial \widehat{w}}{\partial y} = \left[1 + \frac{1}{2}n_g(1+\omega)\frac{w^* M_m \widehat{w}}{M_w\rho_m}\right]\frac{d\delta}{d\tau} \approx \frac{d\delta}{d\tau}. \tag{8.8}$$

8.1. Finite difference scheme. From now on, for the sake of simplicity, we write all the notation without the hat symbol.

Consider (8.1) and (8.6). We note that the diffusion coefficients are $K_i/(1+\omega)^2 \approx 10^{11}$, where $i = s, w$ and $\delta \approx 1$. For this reason we use an implicit numerical scheme—which is stable, monotone, and even uses large time steps—to solve our system.

8.1.1. Full speed case. We assume that $j = 1, \dots, J$ (space index) and $n = 1, \dots, N$ (time index), and we indicate $w(x,t)$ and $s(x,t)$ by the numerical approximations W_i^n and S_j^n. We discretize (8.3) by

$$\frac{\delta^{n+1} - \delta^n}{\Delta\tau} = -\frac{\Omega_s}{\delta^n}\frac{3S_J^n - 4S_{J-1}^n + S_{J-2}^n}{2\Delta y}, \tag{8.9}$$

and (8.1) by

$$\frac{S_j^{n+1} - S_j^n}{\Delta\tau} = y_j\frac{\dot{\delta}^n}{\delta^n}\frac{S_{j+1}^n - S_{j-1}^n}{2\Delta y} + \frac{K_s}{(1+\omega)^2(\delta^{n+1})^2}\frac{S_{j+1}^{n+1} - 2S_j^{n+1} + S_{j-1}^{n+1}}{(\Delta y)^2}, \tag{8.10}$$

where $y_j = \Delta y(j-1)$. Boundary conditions are given by

$$S_1^n = S_a^n, \quad S_J^n = 0.$$

The first system is explicit. System (8.10) is implicit but driven by a standard tridiagonal matrix, and then it is solved by standard methods.

In the same way we solve problem (8.6) by the scheme

$$\frac{W_j^{n+1} - W_j^n}{\Delta\tau} = y_j\frac{\dot{\delta}^n}{\delta^n}\frac{W_{j+1}^n - W_{j-1}^n}{2\Delta y} + \frac{K_w}{(1+\omega)^2(\delta^{n+1})^2}\frac{W_{j+1}^{n+1} - 2W_j^{n+1} + W_{j-1}^{n+1}}{(\Delta y)^2},$$

with the boundary conditions given by

$$W_1^{n+1} = W_a^n, \quad W_J^{n+1} = \frac{4}{3}W_{J-1}^{n+1} - \frac{1}{3}W_{J-2}^{n+1} + \frac{\Omega_s}{\Omega_w}\left(3S_J^{n+1} - 4S_{J-1}^{n+1} + S_{J-2}^{n+1}\right).$$

This yields another algebraic system, based on a tridiagonal matrix, and hence solvable with a low computational cost.

8.1.2. Reduced speed case. If we are in the reduced speed case, we have different boundary conditions for S; see (8.3) and (8.5). This means that we set

$$\alpha = \frac{t^*}{\sigma^*}\left(\frac{1}{W_0 - W_J^n} - \frac{1}{W_0 - W_1}\right), \tag{8.11}$$

and we first compute

$$\delta^{n+1} = \delta^n + \Delta\tau n_g \frac{M_m s^*}{\rho_m M_s}\alpha S_J^n. \tag{8.12}$$

Now, we again have to solve a system like (8.10). However, to obtain the unknown boundary value S_J^{n+1}, we first observe that (8.12) and (8.5) yield

$$-\Omega_s \frac{1}{\delta(\tau)}\frac{\partial s}{\partial y} = n_g \frac{M_m s^*}{\rho_m M_s}\alpha\left(w\right)s. \tag{8.13}$$

Therefore, we can discretize this equation, preserving the second order accuracy, by the relation

$$-\Omega_s \frac{1}{\delta^{n+1}}\frac{\left(3S_J^{n+1} - 4S_{J-1}^{n+1} + S_{J-2}^{n+1}\right)}{2\Delta y} = n_g \frac{M_m s^*}{\rho_m M_s}\alpha S_J^{n+1}. \tag{8.14}$$

8.2. Initial conditions. Our numerical procedure requires $\delta(0) \neq 0$. If we have to deal with the case $\delta(0) = 0$, then we can obtain a quite reasonable guess of the location of the interface and of the SO_2 distribution in the gypsum at some sufficiently small time Δt proceedings as follows. We distinguish two cases: (a) $s_a(0) > 0$; (b) $s_a(0) = s_a^0 = 0$.

(a) From formulas (7.12), (7.13) we know that $\delta(\Delta t) = \sqrt{2\Omega_s s_a^0 \Delta t}$, $s(\eta, \Delta t) = s_a\left(1 - \frac{\eta}{\Delta t}\right)$.

(b) Here we assume that $\dot{s}_a(0) = c > 0$. Then the slope of the free boundary will be finite and can be obtained from the continuity of s_η, s_ϑ in the origin and the equations

$$-\Omega_s s_\eta(\delta, t) = \dot{\delta}, \quad s_\eta \dot{\delta} + s_\vartheta = 0.$$

For $t \downarrow 0$ and $s_\vartheta \to c$ we derive

$$\dot{\delta}(0) = \sqrt{\Omega_s c}.$$

Hence we may take

$$\delta(\Delta t) \simeq \sqrt{\Omega_s c}\Delta t,$$

$$s(\eta, \Delta t) \simeq c\left(1 - \frac{\eta}{\delta(\Delta t)}\right)\Delta t.$$

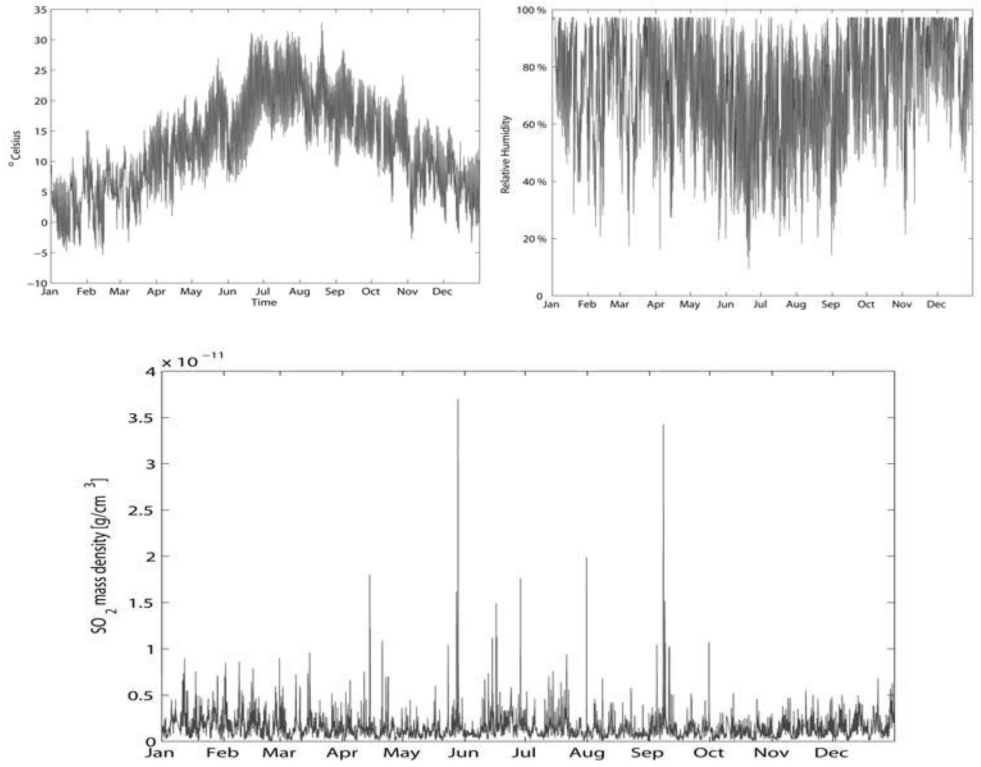


FIG. 1. *Temperature (upper left), relative humidity (upper right), and SO_2 concentration (lower) during* 2006 *at Villa Ada.*

In a realistic situation case (a) is the one of interest. Then we use as initial conditions the solutions of quasi-steady case (7.7) and (7.19) obtained for a time period of one day.

We can note that for each initial condition $s(x, 0)$ and $w(x, 0)$ chosen, these conditions have a completely negligible influence on our solutions after one year of simulations.

9. Numerical simulations. In this section we present some simulations which have been performed using SO_2, humidity, and temperature experimental data detected by Arpalazio, the Roman Regional Authority for air monitoring, every hour during 2006 at Villa Ada in Rome, Italy, a public park with a low level of pollution. Let us also notice that although we simulated the growth of the gypsum crust by our model, we have no experimental data on the real behavior of marble in the same time and under the same conditions. This comparison is beyond the aims of the present work and will be the object of a future work. However, we do have a full agreement between our results and the laboratory tests in [4, 8, 7].

In Figure 1 are shown the temperature (upper left), the relative humidity (upper right), and the distribution of pollutant SO_2 detected at Villa Ada. We get the saturated vapor density (SVD in [g/m^3]) as function of temperature T [°C] using the following relation [10]:

$$SVD(T) = 5.018 + 0.32321T + 8.184710^{-3}T^2 + 3.1243 * 10^{-4}T^3, \tag{9.1}$$

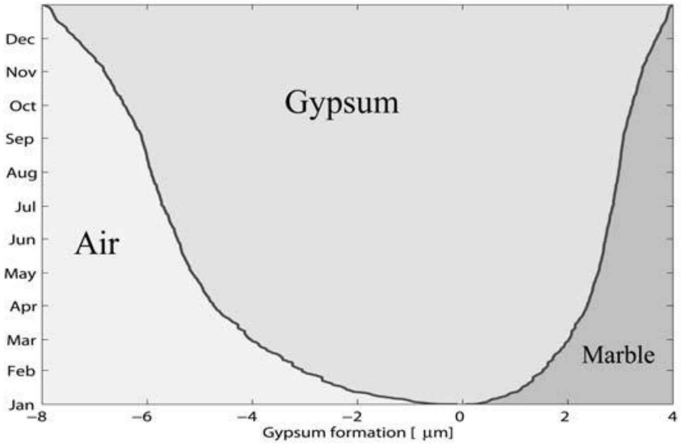


FIG. 2. *Simulated gypsum evolution at Villa Ada.*

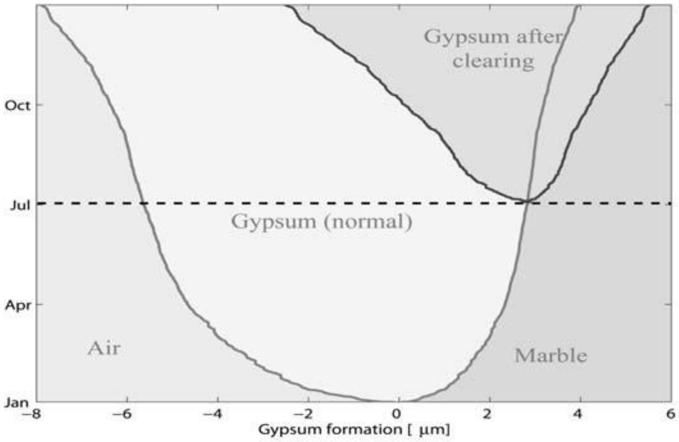


FIG. 3. *Simulated gypsum formation, with and without clearing.*

from which we can obtain the density of vapor in [g/cm^3].

Now, under the conditions detected during 2006 at Villa Ada, we compute marble degradation, starting with $\sigma_0 = 0$. Using our mathematical model with $\Delta t = 10^{-5}$ and $\Delta x = 10^{-1}$, we show in Figure 2 the simulated total thickness of gypsum at the end of 2006. To understand the effect of restoration techniques, let us suppose the gypsum is removed on July 2. We obtain the following simulated gypsum development with and without clearing, shown in Figure 3.

10. Comparison with the quasi-steady solutions. Here, we want to see what happens using quasi-steady solutions, motivated by the fact that the coefficients K_s, K_w are very large. If we assume to have $s(0,t) = s_a =$ const. and $w(0,t) = w_a =$ const., we can observe that the solutions $s(x,t)$ and $w(x,t)$ can be approximated very well by a function linear in x. For this reason, if the fluctuations of boundary values are sufficiently slow, then we can approximate our model by stationary solutions. Here, we want to estimate the error between numerical solutions of the full model and

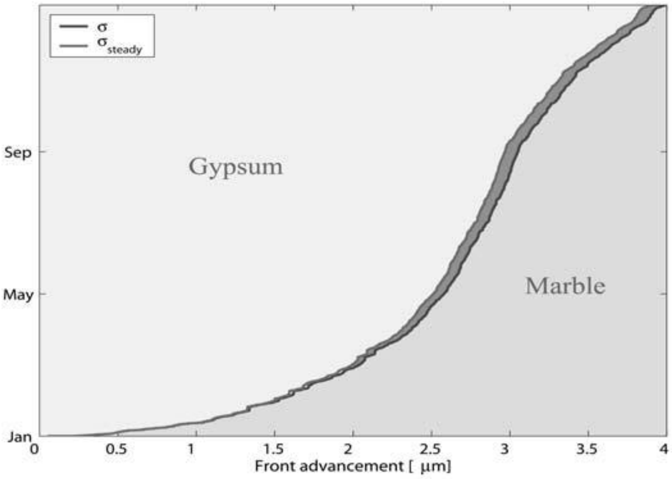


FIG. 4. *Comparison between steady state and numerical solutions at Villa Ada.*

the quasi-steady approximations.

In the full speed case, the system is given by (7.3), (7.7), (7.17), and (7.19). We introduce a new space variable $y = \eta/\delta$, where $y \in [0,1]$, and obtain

$$\dot{\delta} = s_a(\vartheta)\frac{\Omega_s}{\delta(\vartheta)}, \quad s(\eta,\vartheta) = s_a(\vartheta)\,(1-y)\,, \quad w(\eta,\vartheta) = w_a(\vartheta) - \frac{\Omega_s}{\Omega_w}s_a(\vartheta)y.$$

In the reduced speed case, we obtain the analogous equations

$$\dot{\delta}(\tau) = \lambda_1\alpha\frac{s_a(\tau)}{\left(1+\dfrac{\lambda_1}{\Omega_s}\alpha\delta(\tau)\right)},\ s(y,\tau) = s_a(\tau) - \frac{\dot{\delta}(\tau)}{\Omega_s}y\delta(\tau),\ w(y,\tau) = w_a(\tau) - \frac{\dot{\delta}(\tau)}{\Omega_w}y\delta(\tau).$$

To have a quantitative idea of the difference between stationary solutions and numerical solutions, we use the experimental data of Villa Ada. We obtain the behavior of the two fronts of advancement (full model and quasi-stationary solutions), shown on the left of Figure 4. We note that the maximum relative error $\left(\frac{\sigma-\sigma_{steady}}{\sigma}\right)$ obtained is about 2.2%. Therefore this is a case in which it would have been safe to use directly the quasi-steady approximation.

11. Contributions to the front advancement.

11.1. Front advancement as a function of SO$_2$. In the wake of the latter calculation we decided to investigate the influence of SO$_2$ by a factor c, that is to say, we replace $s(0,t)$ by $c \cdot s(0,t)$, where c is a given constant. We used the Villa Ada data and multiplied the SO$_2$ concentration by $c = 1/9$, $1/4$, 1, 2, 4, 6, 9. We can see the result in Figure 5. The results obtained show that the thickness of the front σ, after one year, varies as $\sqrt{c}$. In order to better visualize this fact, in Figure 6 we have reported $\sigma = \sigma(\sqrt{c})$, obtaining a clearly linear graph, consistently with the nature of the quasi-steady approximation.

11.2. Front advancement as a function of time. For the same reason the average growth of σ is close to a $\sqrt{t}$ behavior, strengthening the conclusion reported

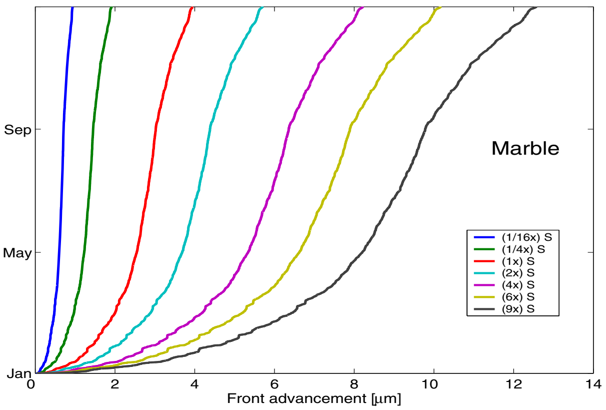


FIG. 5. *Front advancement as a function of SO_2 data.*

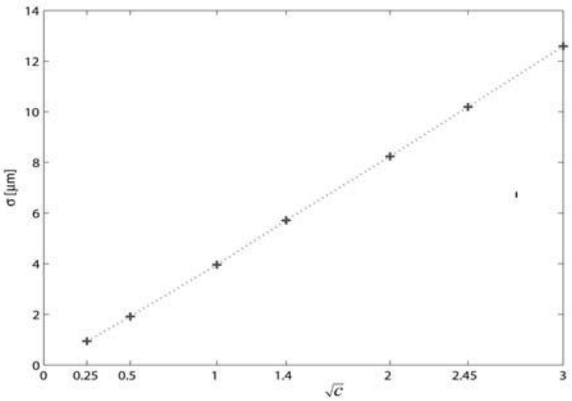


FIG. 6. *Marble wasted after one year for different values of $\sqrt{c}$.*

in [3]. We used the Villa Ada data, and we have repeated the experimental data of the first year (2006) for the next 8 years. This way, we obtained $\sigma(1 \text{ year}) \approx 4\ \mu$m, after 4 years $\sigma(4 \text{ years}) \approx 8.2\ \mu$m, and after 9 years $\sigma(9 \text{ years}) \approx 12.5\ \mu$m. Figure 7 show oscillations around a parabolic behavior.

11.3. Order of accuracy of numerical schemes and extrapolation results. Since we do not have the exact solution of the problem, we give an approximate estimate of the order of the numerical method using the following standard relations:

1. accuracy in L^1 norm:

$$\gamma_1 = \log_2\left(\frac{\|u(h) - u(h/2)\|_1}{\|u(h/2) - u(h/4)\|_1}\right); \tag{11.1}$$

2. accuracy in L^∞ norm:

$$\gamma_\infty = \log_2\left(\frac{\max_x |u(h) - u(h/2)|}{\max_x |u(h/2) - u(h/4)|}\right). \tag{11.2}$$

Using the simulations of the Villa Ada case, with $h = 0.1$, we obtain the results shown in Table 3, which are in good agreement with the formal truncation error of the implicit

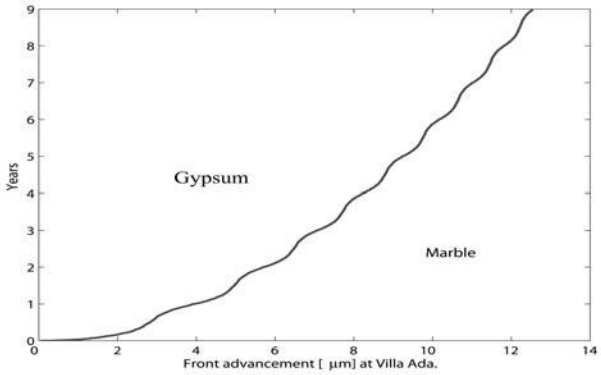


FIG. 7. *Front advancement in* 9 *years at Villa Ada.*

TABLE 3
Approximate order of accuracy of the numerical schemes.

Norm	γ_s	γ_w
L^1	0.89	0.84
L^∞	0.90	0.84

scheme, which is just order 1. Finally further consideration can be done regarding σ obtained after 1 year at Villa Ada. Using $h^* = 0.1$ we found $\sigma_{h^*} = 3.960$ μm, with $h^*/2$, $\sigma_{h^*/2} = 3.971$ μm, and with $h^*/4$, $\sigma_{h^*/4} = 3.977$ μm. Making an extrapolation as $h \to 0$, we obtained an estimate for the real value, namely, $\sigma_{exact} = 3.983$. This way we can estimate that our numerical relative error, in terms of the gypsum front, is about 0.5%.

12. Conclusions.

We have proposed a quantitative model to predict the growth of the gypsum crust on marble stones by using environmental data, i.e., temperature, relative humidity, and pollution concentration. For this model we have identified a quasi-steady asymptotic regime, in good agreement with laboratory data [4, 8, 7] and numerical simulations. Thanks to this model we are now able to quantify how the influence of the evolution of local conditions, or even the cleaning of the stone surface, can change the thickness of the crust, and so the total waste of marble. These results will be useful in the optimal design of future conservation strategies.

Appendix. The flow of air.

So far we have used the approximation (2.5), which has yet to be justified, studying the flow of air through gypsum. In this appendix, we use directly the Lagrangian coordinates, introduced in section 6, to deal with the air pressure $p(x,t) = P(\xi,t)$. In isothermal conditions the air density $\rho_a(\xi,t)$ is a known function of pressure that we can linearize around a reference pressure P_0 (typically 1 Atm.):

$$\rho_a = \rho_0 + \lambda(P - P_0) \tag{A.1}$$

($\rho_0 = \rho_a(P_0)$). The *volumetric* flow of air relative to gypsum can be described by Darcy's law,

$$n_g v_a^R = -k_a \frac{1}{1+\omega}\frac{\partial P}{\partial \xi}, \tag{A.2}$$

where v_a^R is the average molecular velocity in the Lagrangian reference frame, and for hydraulic conductivity of air we take the linear approximation $k_a = k_0 + \chi(P - P_0)$.

In this same frame the air mass balance can be written as

$$\frac{\partial \rho_a}{\partial t} + \frac{1}{1+\omega}\frac{\partial}{\partial \xi}(v_a^R \rho_a) = 0, \tag{A.3}$$

yielding the equation governing the evolution of pressure

$$\begin{aligned} &\lambda \frac{\partial P}{\partial t} - \frac{1}{(1+\omega)^2}[k_0\rho_0 + (k_0\lambda + \rho_0\chi)(P - P_0)]\frac{\partial^2 P}{\partial \xi^2} \\ &- \frac{1}{(1+\omega)^2}\{\lambda[k_0 + \chi(P - P_0)] + \chi[\rho_0 + \lambda(P - P_0)]\}\Big(\frac{\partial P}{\partial \xi}\Big)^2 = 0, \end{aligned} \tag{A.4}$$

where we have neglected the terms with $(P - P_0)^2$, consistently with the linearization already performed.

In conditions of still external air the value of P at $\xi = 0$ can be set equal to the value in the atmosphere (the wind can alter the air pressure at the gypsum surface):

$$P(0, t) = P_a(t). \tag{A.5}$$

In order to write down the mass balance of air on the reaction front $\xi = \sigma(t)$, we suppose that marble pores are filled with air having a prescribed density ρ^0 (for instance, $\rho^0 = \rho_0$), and we impose that the air flux supplies the amount of air $(n_g\rho_a - n_m\rho^0)\dot{\sigma}$ per unit time and unit surface of the front (n_m is the marble porosity). Performing the usual linearization we obtain

$$\begin{aligned} &-\frac{1}{1+\omega}[k_0\rho_0 + (k_0\lambda + \rho_0\chi)(P - P_0)]\frac{\partial P}{\partial \xi} \\ &= n_g[\rho_0 + \lambda(P - P_0)](1+\omega)\dot{\sigma} - n_m\rho^0\dot{\sigma}. \end{aligned} \tag{A.6}$$

The justification of (2.5) now comes from rescaling, as we shall see next.

Rescaling the air flow problem. Rescaling P, following the indications in section 7, we get $\widehat{P}(\eta, \vartheta) = P(\xi, t)/P_0$. In the new variables (A.4) is written as

$$\frac{\partial \widehat{P}}{\partial \vartheta} - \frac{[1 + (A+B)(\widehat{P} - 1)]}{(1+\omega)^2}K_a\frac{\partial^2 \widehat{P}}{\partial \eta^2} = \frac{[A + B + 2AB(\widehat{P} - 1)]}{(1+\omega)^2}K_a\Big(\frac{\partial \widehat{P}}{\partial \eta}\Big)^2, \tag{A.7}$$

containing the nondimensional coefficients $A = P_0\frac{\lambda}{\rho_0}$, $B = P_0\frac{\chi}{k_0}$, $K_a = \frac{t^* k_0\rho_0}{\lambda\sigma^{*2}}$. We remark that $t_0 = k_0\rho_0$ has the dimension of time and that $\lambda^{-1/2} = v_\lambda$ is the reference value of the speed of sound in air (340 m/sec). Thus the constant K_a can be written as

$$K_a = \frac{v_\lambda^2}{v^{*2}}\frac{t_0}{t^*}, \tag{A.8}$$

with $v^* = \frac{\sigma^*}{t^*}$ representing a typical mean velocity of the reaction front.

Let us derive the nondimensional version of the balance equation (A.6), noting that $\dot{\sigma} = \frac{\sigma^*}{t^*}\frac{d\delta}{d\vartheta} = v^*\frac{d\delta}{d\vartheta}$ and that $\frac{P_0 t_0}{\rho_0 v^* \sigma^*} = AK_a$:

$$-\frac{[1 + (A+B)(\widehat{P} - 1)]}{(1+\omega)^2}AK_a\frac{\partial \widehat{P}}{\partial \eta} = n_g\frac{d\delta}{d\vartheta}\Big[1 + A(\widehat{P} - 1) - \frac{1}{1+\omega}\frac{n_m}{n_g}\frac{\rho^0}{\rho_0}\Big]. \tag{A.9}$$

Since $A \simeq 1$, if $K_a \gg 1$ the air motion is quasi steady and $Q = \frac{\partial \widehat{P}}{\partial \eta}$ obeys

$$[1 + (A + B)(\widehat{P} - 1)]\frac{\partial Q}{\partial \eta} + [A + B + 2AB(\widehat{P} - 1)]Q^2 = 0,$$

with the condition $Q(\delta(\vartheta)) \simeq 0$, implying $\frac{\partial \widehat{P}}{\partial \eta} \simeq 0$ throughout the gypsum.

Therefore, if $K_a \gg 1$, the pressure field is flat. To check this property, we observe that $v_\lambda^2 \simeq (3.4 \cdot 10^4 \text{cm} \cdot \text{sec}^{-1})^2$, $v^{*2} = \left(\frac{2 \cdot 10^{-3}\text{cm}}{3.15 \cdot 10^7 \text{sec}}\right)^2$, and hence $\frac{v_\lambda^2}{v^{*2}} \simeq 2.5 \cdot 10^{28}$, $t_0 \simeq 10^{-13}$sec since $k_0 \simeq 10^{-10}\text{g}^{-1}\text{cm}^3\text{sec}$ and $\rho_0 \simeq 10^{-3}\text{g}\cdot\text{cm}^{-3}$, and thus $\frac{t_0}{t^*} \simeq \frac{10^{-20}}{3}$. So finally $K_a \simeq 10^8$ and formula (2.5) is largely justified.

Acknowledgments. The authors thank the Roman Regional Authority for environment, Arpalazio (http://www.arpalazio.it/), Agenzia Regionale per la Protezione Ambientale del Lazio, for all the in situ data used in this paper. They also wish to thank Deepika Ghai for the work done on this problem during her summer internship in 2006 at Istituto per le Applicazioni del Calcolo in Rome.

REFERENCES

[1] G. Alì, V. Furuholt, R. Natalini, and I. Torcicollo, *A mathematical model of sulfite chemical aggression of limestones with high permeability. Part* I. *Modeling and qualitative analysis*, Transp. Porous Media, 69 (2007), pp. 109–122.

[2] G. G. Amoroso and V. Fassina, *Stone Decay and Conservation: Atmospheric Pollution, Cleaning, Consolidation and Protection*, Elsevier Science Publishers, Amsterdam, 1983.

[3] D. Aregba-Driollet, F. Diele, and R. Natalini, *A mathematical model for the sulphur dioxide aggression to calcium carbonate stones: Numerical approximation and asymptotic analysis*, SIAM J. Appl. Math., 64 (2004), pp. 1636–1667.

[4] E. Borrelli, C. Giavarini, M. Incitti, M. L. Santarelli, and R. Natalini, *A material model for the evolution of gypsum crusts: Numerical and experimental results*, in Proceedings of the 10th International Congress on Deterioration and Conservation of Stone, Stockholm, Sweden, 2004, Vol. 1, D. Kwiatkowski and R. Löfvendahl, eds., ICOMOS, pp. 35–42.

[5] E. A. Charola, *Acidic deposition on stone*, US/ICOMOS Sci. J., 3 (2001), pp. 19–58.

[6] J. D. Evans and J. R. King, *Asymptotic results of the Stefan problem with kinetic undercooling*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 449–473.

[7] K. L. Gauri and J. K. Bandyopadhyay, *Carbonate Stone, Chemical Behavior, Durability, and Conservation*, John Wiley & Sons, New York, 1999.

[8] C. Giavarini, M. L. Santarelli, R. Natalini, and F. Freddi, *A nonlinear model of sulfation of porous stones: Numerical simulations and preliminary laboratory assessments*, J. Cultural Heritage, 9 (2008), pp. 14–22.

[9] F. R. Guarguaglini and R. Natalini, *Fast reaction limit and large time behavior of solutions to a nonlinear model of sulfation phenomena*, Comm. Partial Differential Equations, 32 (2007), pp. 163–189.

[10] *HyperPhysics website*, http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html, Georgia State University.

[11] J. L. Pérez Bernal and M. A. Bello, *Modeling sulfur dioxide deposition on calcium carbonate*, Ind. Eng. Chem. Res., 42 (2003), pp. 1028–1034.

SIAM J. Appl. Math.
Vol. 69, No. 1, pp. 169–196

NEURAL ASSOCIATIVE MEMORY AND THE WILLSHAW–PALM PROBABILITY DISTRIBUTION*

ANDREAS KNOBLAUCH†

Abstract. Previous asymptotic analyses of binary neural associative networks of Willshaw or Steinbuch type relied on a binomial approximation of the neurons' dendritic potentials. This approximation has been proven to be good only if the stored patterns are extremely sparse, for example, when the mean number of active units k per pattern vector scales with the logarithm of the vector size n. Recent promising results concerning storage capacity and retrieval efficiency for larger pattern activities $k > \log n$ have been doubted because here the binomial approximation can lead to a massive overestimation of performance. In this work I compute and characterize the exact Willshaw–Palm distribution of the dendritic potentials for hetero-association, auto-association, and fixed and random pattern activity. Comparing the raw and central moments of the Willshaw–Palm distribution to the moments of the corresponding binomial probability reveals that, asymptotically, the binomial approximation becomes exact for almost any sublinear pattern activity, including $k = O(n/\log^2 n)$. This verifies, for large networks, the existence of a wide high-performance parameter range as predicted by the approximative theory.

Key words. neural network, Willshaw model, information retrieval, storage capacity, fault tolerance

AMS subject classifications. 92B20, 68T10, 60C05

DOI. 10.1137/070700012

1. Introduction.

Associative memories are systems that contain information about a finite set of associations between pattern vector pairs $\{(\mathbf{u}^\mu \mapsto \mathbf{v}^\mu) : \mu = 1, \ldots, M\}$, where $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ are called *address* and *content* patterns, respectively [28]. Given a possibly noisy address pattern $\tilde{\mathbf{u}}$ the problem is to find a content pattern $\mathbf{v}^\mu$ for which the corresponding address pattern $\mathbf{u}^\mu$ is most similar to $\tilde{\mathbf{u}}$. This is a variant of the *best match problem* in [31], and efficient solutions have widespread applications, including object recognition and information retrieval [28, 36, 40, 3, 13, 20, 32, 42].

In *neural network implementations* the information about the associations is stored in the synaptic connectivity of one or more neuron populations [46, 16, 17, 37]. Besides the potential for technical applications, neural associative memories also play an important role in many *brain theories* (e.g., [14, 30, 5, 35, 16, 17, 11, 12, 27, 10, 15]), where the patterns correspond to attractors in the brain's neuronal state space.

One of the most efficient networks is the so-called Willshaw or Steinbuch model with binary neurons and synapses [44, 46, 34, 33, 8, 43]. In particular, it has been shown that the Willshaw model has a very high asymptotic storage capacity of $C = 0.7$ bits per synapse which exceeds the capacity of most alternative models. For example, the original Hopfield model achieves only $C = 0.14$ bits per synapse [16, 1, 2]. In general the classical work points out that high capacities can be obtained only if the stored patterns are extremely sparse, for example, when the mean number of active units k per pattern vector scales logarithmic with the vector size n.

For a number of reasons, a regime of larger pattern activity with $k/\log n \to \infty$ has recently gained increased attention: First, logarithmic $k = \log n$ is simply too sparse

*Received by the editors August 13, 2007; accepted for publication (in revised form) May 28, 2008; published electronically September 8, 2008.
http://www.siam.org/journals/siap/69-1/70001.html

†Honda Research Institute Europe, Carl-Legien-Strasse 30, D-63073 Offenbach/Main, Germany (andreas.knoblauch@honda-ri.de).

for many applications of distributed representations [41, 45, 42]. Second, activity patterns with extremely sparse activity $k \sim \log n$ appear inconsistent with neurophysiology because they are difficult to stabilize in a noisy regime where neurons have high rates of spontaneous activity [29]. Third, it has been argued that $k/\log n \to \infty$ can actually lead to a massive increase in storage capacity and retrieval efficiency if the network structures are adequately compressed ([22]; see also [18, 19, 20]). Fourth, $k/\log n \to \infty$ allows an efficient inhibitory implementation of the Willshaw model which implies new interpretations for inhibitory circuits in the brain [22, 24].

However, the viability of this regime with moderately sparse patterns, $k/\log n \to \infty$, has been doubted. On the one hand, here the established theories on Willshaw- or Steinbuch-type networks with fixed connectivity structure predict only a very low performance, for example, zero storage capacity per synapse, such that both technical applicability and biological relevance seem unlikely. On the other hand, the extended theory considering structural changes and inhibitory implementations predicts high performance but, similar to the established theories, relied on a *binomial approximation* of the neurons' dendritic potentials (e.g., [46, 34, 37, 33, 4, 43, 20]). This approximation may be inaccurate for large pattern activities $k \gg \log n$ and thus the corresponding high-performance regime illusory. Indeed, the convergence of the binomial approximation to the true potential distribution and thus the asymptotic correctness of the theory has been demonstrated only for some special cases involving very sparse activity patterns, where a binary pattern vector of n neurons contains only $k = \log n$ or $k \leq n^{1/3}$ active units [34, 38]. Another analysis showed that the binomial approximation becomes very inaccurate for linear $k \sim n$ [19, 21]. However, it remained unclear for precisely which $k(n)$ the binomial approximation converges to the true potential distribution.

In this work I have solved this problem. Section 2 gives an overview of the Willshaw model and the analysis employing the binomial approximation of the dendritic potentials. Section 3 then defines and computes the exact Willshaw–Palm distribution of the dendritic potentials which can be used to determine exact retrieval error probabilities and storage capacity. Section 4 characterizes the Willshaw–Palm probability by computing the raw and central moments. Finally, section 5 compares the Willshaw–Palm probability to the binomial probability and determines asymptotic conditions when the two probability distributions become identical.

2. Binary associative networks.

2.1. Learning and retrieving patterns.

An attractive model of neural associative memory both for biological modeling and applications is the so-called Willshaw or Steinbuch model with binary neurons and synapses [46, 44, 34, 33, 37, 7, 4, 43, 20] illustrated in Figure 2.1. Each address pattern $\mathbf{u}^\mu$ is a binary vector of length m containing k one-entries and $m-k$ zero-entries. Similarly, each content pattern $\mathbf{v}^\mu$ is a binary vector of length n containing l one-entries and $n-l$ zero-entries. Typically, the patterns are sparse, i.e., $k \ll m$ and $l \ll n$. For our analysis of storage capacity we will further assume that each pattern is randomly drawn from the sets of the $\binom{m}{k}$ potential address patterns and the $\binom{n}{l}$ potential content patterns.

The M pattern pairs are stored *hetero-associatively* in a binary *memory matrix* $\mathbf{A} \in \{0,1\}^{m \times n}$ with

$$A_{ij} = \min\left(1, \tilde{A}_{ij} + \sum_{\mu=1}^{M} u_i^\mu \cdot v_j^\mu\right) \in \{0,1\}, \tag{2.1}$$

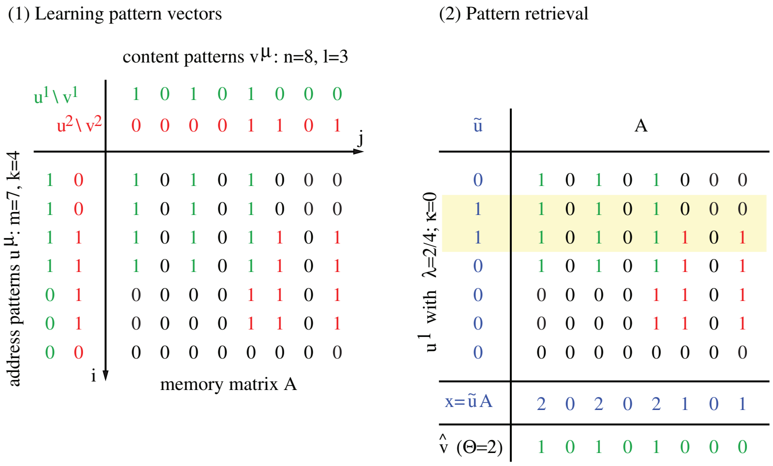


FIG. 2.1. *Example of the binary Willshaw associative memory for hetero-association. Left: During learning, M associations between address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ are stored in the binary memory matrix $\mathbf{A}$ representing binary synaptic weights of the connection from neuron population u to v. Initially all synapses are inactive ($\tilde{p}_1 = 0$). During learning of pattern associations, the synapses are activated according to Hebbian coincidence learning (equation (2.1)). Right: For retrieval an address pattern $\tilde{\mathbf{u}}$ is propagated through the network. Vector-matrix multiplication yields the dendritic potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$. To obtain the retrieval result $\hat{\mathbf{v}}$ (here equal to $\mathbf{v}^1$) a threshold Θ is applied. For pattern part retrieval with $\tilde{\mathbf{u}} \subseteq \mathbf{u}^\mu$ we can simply choose the Willshaw threshold $\Theta = |\tilde{\mathbf{u}}|$. Then the retrieval output is a superset of the original pattern, $\hat{\mathbf{v}} \supseteq \mathbf{v}^\mu$, which means $\hat{\mathbf{v}}$ contains no miss-errors.*

where $\tilde{\mathbf{A}}$ is a binary noise matrix with each component being active independently with probability $\tilde{p}_1$.

The *neural interpretation* is that of two neuron populations, an address population u consisting of m neurons and a content population v consisting of n neurons. The patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ describe the activity states of the two populations at time μ, and A_{ij} is the strength of the Hebbian learned synaptic connection from neuron u_i to neuron v_j. Positive $\tilde{p}_1$ can be used to model noisy synaptic potentiation (e.g., the synapses that are already active before learning starts), noisy synaptic transmission, or incomplete connectivity [22, 23, 24].

Besides the feed-forward interpretation, the Willshaw model can also be used to model *auto-association* or pattern completion where address population content population are identical, $u = v$, and consequently also $\mathbf{u}^\mu = \mathbf{v}^\mu$. Here the memory matrix $\mathbf{A}$ describes the recurrent synaptic connectivity within the neuron population.

For independently generated random patterns, there is a simple relation between the number M of stored associations and the so-called *memory load* p_1 defined as the fraction of one-entries in the memory matrix. The probability that a synapse is *not* activated by the association of one pattern pair is $1 - kl/mn$; therefore, after learning M pattern associations,

$$p_1 = 1 - (1 - \tilde{p}_1)\left(1 - \frac{kl}{mn}\right)^M \geq \tilde{p}_1, \tag{2.2}$$

$$M = \frac{\ln \frac{1-p_1}{1-\tilde{p}_1}}{\ln(1 - kl/mn)} \approx -\frac{mn}{kl} \ln \frac{1 - p_1}{1 - \tilde{p}_1}, \tag{2.3}$$

where the approximation is valid for $kl \ll mn$. As we will see, the memory load p_1

will play an important role both for the exact analysis of the Willshaw model and for the binomial approximative analysis.

After learning, the stored information can be retrieved by applying an address pattern $\tilde{\mathbf{u}}$. Vector-matrix multiplication yields the dendritic potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$ of the content neurons, and imposing a threshold Θ gives the (one-step) retrieval result $\hat{\mathbf{v}}$,

$$\hat{v}_j = \begin{cases} 1, & x_j = (\sum_{i=1}^{m} \tilde{u}_i A_{ij}) \geq \Theta, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Choosing $\Theta = z := \sum_{i=1}^{m} \tilde{u}_i$ will be referred to as the *Willshaw threshold* and plays an important role both for more realistic *spiking* neuron networks [26, 19] and also for pattern part retrieval with $\tilde{\mathbf{u}} \subseteq \mathbf{u}^\mu$ as analyzed in section 2.3.

2.2. Retrieval errors and storage capacity. We have retrieval errors if the retrieval result $\hat{\mathbf{v}}^\mu$ is not identical to the originally learned pattern $\mathbf{v}^\mu$. For a closer analysis we can divide the neurons of the content population into two groups: The *lo-units*, which correspond to the $n - l$ zero-entries of $\mathbf{v}^\mu$, and the *hi-units*, which correspond to the l one-entries of $\mathbf{v}^\mu$. For an error-free retrieval result $\hat{\mathbf{v}}^\mu$ the potentials $\mathbf{x}$ of lo- and hi-units must be separable; i.e., the largest potential of a lo-unit must be smaller than the smallest potential of a hi-unit. If the two potential distributions have overlap, two kinds of retrieval errors can occur. An *add-error* occurs if the potential of a lo-unit is above threshold Θ, and a *miss-error* occurs if the potential of a hi-unit is below threshold. If the probability distribution of a lo-unit i is known, we can compute the probability p_{01} of an *add-error*. Similarly, for a hi-unit j we can compute the probability p_{10} of a *miss-error*. With $z = |\tilde{\mathbf{u}}|$ being the activity of the address pattern, we have

$$p_{01} = \mathrm{pr}(\hat{v}_i = 1 | v_i^\mu = 0) = \sum_{x=\Theta}^{z} \mathrm{pr}[x_i = x], \tag{2.5}$$

$$p_{10} = \mathrm{pr}(\hat{v}_j = 0 | v_j^\mu = 1) = \sum_{x=0}^{\Theta-1} \mathrm{pr}[x_j = x]. \tag{2.6}$$

Thus, the expected *Hamming distance* $h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) := \sum_{j=1}^{n} |v_j^\mu - \hat{v}_j^\mu|$ between learned and retrieved patterns is

$$Eh(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = (n - l)p_{01} + lp_{10}. \tag{2.7}$$

To enforce retrieval quality we bound the expected Hamming distance to be no more than a fraction ϵ of the content pattern activity l. Thus, we require

$$(n - l)p_{01} + lp_{10} \leq \epsilon l, \tag{2.8}$$

where retrieval quality parameter ϵ is typically a small positive constant (e.g., $\epsilon = 0.01$). Because the minimal Hamming distance (optimizing Θ) is obviously increasing with M, we can finally define the *pattern capacity* M_ϵ,

$$M_\epsilon := \max\{M : (n - l)p_{01} + lp_{10} \leq \epsilon l\}, \tag{2.9}$$

being the maximal number of storable pattern associations fulfilling the retrieval quality requirement (2.8). Considering the Shannon information of individual content patterns, we get the normalized *network storage capacity* in bits per synapse,

$$C_\epsilon := \frac{M_\epsilon T(\mathbf{v}^\mu; \hat{\mathbf{v}}^\mu)}{mn}, \tag{2.10}$$

where $T(\mathbf{v}^\mu;\hat{\mathbf{v}}^\mu)$ is the transinformation (or mutual information) between learned and retrieved content patterns [9]. From the network capacity we can derive further performance measures such as *information capacity* C^I and *synaptic capacity* C^S, making use of the compressibility of the memory matrix for a memory load $p_1 \neq 0.5$ (see section 2.4; for more details see [18, 19, 20, 22]).

2.3. Sketch of the binomial approximative analysis for random patterns. The approximative analysis of the Willshaw model relies on the assumption that the one-entries in the memory matrix are generated independently of each other. Although obviously not true for distributed patterns, this assumptions leads to seminal insights into the Willshaw model and, at least for certain parameter ranges, quite good approximations of the actual storage capacity (see section 3 and [22]).

Let us again assume that the retrieval address pattern $\tilde{u}$ contains $c = \lambda k$ correct and $f = \kappa k$ false one-entries of address pattern u^μ previously used for learning $(0 < \lambda \leq 1,\ \kappa \geq 0)$. Assuming $\mathrm{pr}[A_{ij} = 1] = p_1$ independently of i, j, the dendritic potentials x_{lo} of a lo-unit and x_{hi} of a hi-unit are binomially distributed (equation (A.2)),

$$\mathrm{pr}[x_{\mathrm{lo}} = x] = p_B(x; c+f, p_1), \quad x = 0, 1, \ldots, c+f, \tag{2.11}$$

$$\mathrm{pr}[x_{\mathrm{hi}} = x] = p_B(x-c; f, p_1), \quad x = c, c+1, \ldots, c+f. \tag{2.12}$$

For purposes of clarity, in the following we restrict the analysis to the case of *pattern part retrieval* where the address pattern contains no add-noise, that is, $f = 0$. For the general analysis see [43]. Here one can apply the Willshaw threshold $\Theta = c$, which will limit the retrieval errors to add-noise. Thus, the retrieval error probabilities are

$$p_{01} = p(\hat{v}_i = 1 | v_i^\mu = 0) \approx {p_1}^{\lambda k} \tag{2.13}$$

and $p_{10} = 0$. To enforce retrieval quality as described above (see (2.8)) we have to bound the error probability p_{01} by $p_{01\epsilon}$,

$$p_{01} \leq p_{01\epsilon} := \frac{\epsilon l}{n-l}. \tag{2.14}$$

The number of patterns that can be stored is limited to the point where $p_{01} = p_{01\epsilon}$ or, equivalently, where the memory load reaches

$$p_{1\epsilon} \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda \cdot k}} \qquad \left(\Leftrightarrow k \approx \frac{\mathrm{ld}\,\frac{\epsilon l}{n-l}}{\lambda\,\mathrm{ld}\,p_{1\epsilon}}\right). \tag{2.15}$$

From (2.3) we obtain the maximal number of stored patterns or *pattern capacity*

$$M_\epsilon \approx -\lambda^2 \cdot (\mathrm{ld}\,p_{1\epsilon})^2 \cdot \ln\frac{1-p_{1\epsilon}}{1-\tilde{p}_1} \cdot \frac{k}{l} \cdot \frac{mn}{(\mathrm{ld}\,\frac{n-l}{\epsilon \cdot l})^2}. \tag{2.16}$$

With this result we can also estimate the network capacity (equation (2.10))

$$C_\epsilon = \frac{M_\epsilon T(l/n, p_{01\epsilon}, 0)}{m} \approx \lambda \cdot \mathrm{ld}\,p_{1\epsilon} \cdot \ln\frac{1-p_{1\epsilon}}{1-\tilde{p}_1} \cdot \eta, \tag{2.17}$$

where $T(p, p_{01}, p_{10})$ is the transinformation (or mutual information) of a binary channel (see (A.1), [9]) and

$$\eta := \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{-\frac{l}{n}\,\mathrm{ld}\,\frac{\epsilon l}{n-l}} \approx \frac{1}{1 + \frac{\ln \epsilon}{\ln(l/n)}}. \tag{2.18}$$

The approximation is valid for small $\epsilon, l/n \ll 1$ when $T \approx -(l/n)\,\mathrm{ld}(l/n)$: In that case $\eta \to 1$ for large $n \to \infty$. For $p_{1\epsilon} = 0.5$ and $\tilde{p}_1 = 0$ we therefore have $C_\epsilon \to \ln 2 \approx 0.69$ bits per synapse, the asymptotic storage capacity of the Willshaw model [46, 34, 43, 22]. Note that C_ϵ increases by factor $1/(1-\tilde{p}_1)$ if $1-\tilde{p}_1$ is interpreted as network connectivity (i.e., the chance that a potential synapse is actually realized; see [22, 8, 4]).

2.4. The asymptotic regimes of sparse and dense potentiation. The main conclusions from the binomial approximative analysis are that a very high storage capacity of almost 0.7 bits per synapse can be achieved for sparse patterns with $k \sim \log n$ and memory load $p_1 = 0.5$. Then we can store on the order of $M \sim mn/(\log n)^2$ pattern associations with high retrieval quality. From (2.15), (2.17) it is easy to see that asymptotically

$$C_\epsilon > 0 \Leftrightarrow k \sim \log n \Leftrightarrow 0 < p_{1\epsilon} < 1. \tag{2.19}$$

Thus, the analysis suggests that neural associative memory is efficient ($C_\epsilon > 0$) only for logarithmically sparse patterns. For sublogarithmic sparse patterns with $k/\log n \to 0$ we have $p_{1\epsilon} \to 0$, and for supralogarithmic sparseness with $k/\log n \to \infty$ we have $p_{1\epsilon} \to 1$, both cases implying vanishing network storage capacity $C_\epsilon \to 0$. These results bear importance for both technical and biological applications, in particular with respect to the sparseness of postulated Hebbian cell assemblies in the brain [14, 5, 35]. In the following we will refer to the three cases $p_{1\epsilon} \to 0/c/1$ as sparse, balanced, and dense synaptic potentiation, respectively.

I have argued elsewhere that these conclusions may be biased by the definition of network storage capacity, and that alternative definitions of storage capacity considering the compressibility of the network lead to different conclusions [18, 19, 20, 22]. For example, in technical implementations of the Willshaw model the memory matrix can be compressed for $p_1 \to 0/1$ and the storage capacity improves by factor $I(p_1) := -p_1\,\mathrm{ld}\,p_1 - (1-p_1)\,\mathrm{ld}(1-p_1)$. Similar arguments hold for biological networks where “compression” could be realized by synaptic pruning and structural plasticity (see [22] for more details). This has led to the definition of information capacity $C_\epsilon^I := C_\epsilon/I(p_{1\epsilon})$ and synaptic capacity $C_\epsilon^S := C_\epsilon/\min(p_{1\epsilon}, 1-p_{1\epsilon})$. Interestingly, and in contrast to network capacity C_ϵ, optimizing C_ϵ^I and C_ϵ^S reveals highest capacities for $p_{1\epsilon} \to 0$ and $p_{1\epsilon} \to 1$. Here, presuming the validity of the binomial theory, technical implementations could fully exploit the physical memory by storing $C_\epsilon^I \to 1$ bit information per memory bit. Similarly, biological networks could improve storage capacity to arbitrary large values $C_\epsilon^S \sim \log n \to \infty$ bits per synapse. By these results, the regimes with ultrasparse and moderately sparse patterns (or cell assemblies) have gained increased attention. However, the convergence of the binomial approximations towards the exact values is questionable since this has been strictly proven only for some special conditions including $k \sim \log n$ [34, 38]. In particular, for dense potentiation with $p_{0\epsilon} = 1 - p_{1\epsilon} \to 0$, supralogarithmic sparseness, $k/\log n \to \infty$, and

$$p_{1\epsilon} = \left(\frac{\epsilon l}{n-l}\right)^{1/\lambda k} = e^{\frac{\ln(\epsilon l/(n-l))}{\lambda k}} \approx 1 - \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k}, \tag{2.20}$$

numerical simulations of the Willshaw model reveal that the real capacities can be massively overestimated by the binomial approximative analysis [22]. Therefore, in the following we conduct an exact analysis of the Willshaw model based on the exact potential distributions, and investigate conditions when the binomial probability distribution becomes a good approximation of the Willshaw–Palm distribution.

3. The Willshaw–Palm distribution of the dendritic potentials.

For an exact analysis of the Willshaw model we have to compute the distribution of the neurons' dendritic potentials, i.e., the probability $\mathrm{pr}[X = x]$ that the potential X of a given hi- or lo-unit equals a certain value x (see (2.5), (2.6)). This probability distribution is also called *Willshaw–Palm distribution* for random pattern associations and random retrieval address pattern. In the following more formal definition we take into account different ways to generate random patterns.

Definition 3.1 (Willshaw–Palm probability). *Let A be the memory matrix of a Willshaw associative memory after learning M random pattern associations and with synaptic noise $\tilde{p}_1$ as described in section* 2.1. *The associations are between address patterns u^μ with size m and mean activity k, and content patterns v^μ with size n and mean activity l ($\mu = 1, 2, \ldots, M$). Further let $\tilde{u}$ be a binary random address pattern with activity $z = |\tilde{u}|$. Then we define the* Willshaw–Palm probability *as the probability* $\mathrm{pr}[(\tilde{u}A)_j = x]$ *that a given content neuron v_j has potential x when retrieving with $\tilde{u}$. We distinguish between four relevant versions of the Willshaw–Palm probability depending on the generation of the random patterns:*

1. $p_{\mathrm{Ph}}(x; k, l, m, n, M, \tilde{p}_1, z)$ *for* fixed *address activity and* hetero-association.
2. $p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, \sigma)$ *for* fixed *address activity and* auto-association.
3. $p_{\mathrm{Wh}}(x; k, l, m, n, M, \tilde{p}_1, z)$ *for* random *address activity and* hetero-association.
4. $p_{\mathrm{Wa}}(x; k, n, M, \tilde{p}_1, z, \sigma)$ *for* random *address activity and* auto-association.

Auto-association *means that address patterns and content patterns are identical, $u^\mu = v^\mu$.* Fixed *address activity means that each address pattern has exactly k active units.* Random *address activity means that a component of an address pattern is active, $u_i^\mu = 1$, with probability k/m independently of other components. For the hetero-associative cases, the* content *patterns can have either fixed activity l or random activity with mean l. The auto-associative cases require an additional parameter $\sigma := \mathrm{pr}[j \in \tilde{u}]$ denoting the probability that neuron j is among the z active address units.*

We sometimes denote p_{W} briefly as the *Willshaw probability* since p_{Wh} has first been determined by Buckingham and Willshaw [7, 6]. Similarly, we denote p_{P} briefly as the *Palm probability* since some special cases of p_{Ph} have first been determined by Palm [34]. Note that the difference between the two variants is that the Palm model has address patterns with fixed activity and the Willshaw model has address patterns with fixed mean.

Theorem 3.2. *The four Willshaw–Palm probabilities p_{Ph}, p_{Pa}, p_{Wh}, p_{Wa} are given by* (3.22), (3.34), (3.39), (3.41), *respectively.*

The proof of the theorem follows in the next four subsections, each determining one version of the Willshaw–Palm probability and the corresponding retrieval error probabilities.

3.1. Fixed pattern activity and hetero-association.

Here we will determine the Willshaw–Palm probability $p_{\mathrm{Ph}}(x; k, l, m, n, M, z)$ of Definition 3.1. For brevity we identify patterns with sets of one-entries, e.g., $\mathbf{u} = 011001$ is identified with the index set $\mathbf{u} = \{2, 3, 6\}$. Generalizing Palm's definition of a predicate or condition C (see appendix 1 in [34]) for index sets Y, N ("yes!" and "no!") let

$$C(Y, N, j) := [\forall i \in Y : A_{ij} = 1] \cap [\forall i \in N : A_{ij} = 0]; \tag{3.1}$$

i.e., condition $C(Y, N, j)$ means that content neuron j receives inputs from the subset Y of address pattern $\tilde{\mathbf{u}}$, but no input from subset N. We further assume that Y and N are disjunct random sets unrelated to the M stored pattern pairs. For Y equal to a further $(M+1)$th address pattern, i.e., $Y = \mathbf{u}^{M+1}$, the condition $C(Y, \emptyset, j)$ would coincide with the definition of C in the appendix of [34]. Then C would be equivalent to the occurrence of an add-error at lo-unit j for retrieval with the noise-free address pattern $\tilde{\mathbf{u}} = \mathbf{u}^{M+1}$. We first compute the probability that $C(Y, \emptyset, j)$ holds after storing M pattern associations. Contrary to [34] we assume that $Y \subseteq \{1, \dots, m\}$ is an arbitrary *subset* of address units unrelated to the M stored pattern associations.

$$\operatorname{pr}(C(Y,\emptyset,j)) = \operatorname{pr}([\forall i \in Y : A_{ij} = 1]) = 1 - \operatorname{pr}([\exists i \in Y : A_{ij} = 0]) \tag{3.2}$$

$$= 1 - \operatorname{pr}\left(\bigcup_{i \in Y} [A_{ij} = 0] \right) = 1 - \operatorname{pr}\left(\bigcup_{i=1}^{|Y|} [A_{ij} = 0] \right) \tag{3.3}$$

$$= 1 - \sum_{s=1}^{|Y|} (-1)^{s+1} \sum_{1 \le i_1 < \dots < i_s \le |Y|} \operatorname{pr}\left(\bigcap_{h=1}^{s} [A_{i_h j} = 0] \right) \tag{3.4}$$

$$= 1 - \sum_{s=1}^{|Y|} (-1)^{s+1} \binom{|Y|}{s} \operatorname{pr}\left(\bigcap_{i=1}^{s} [A_{ij} = 0] \right). \tag{3.5}$$

For (3.4) we used the formula of Sylvester–Poincaré equation (A.6). Note that for random patterns the probabilities that a given subcolumn of $\mathbf{A}$ has at least one zero-entry (equation (3.3)) or only zero-entries (see (3.5)) depend only on the subcolumn's size, but not on the specific indices. The latter probability is written as

$$\operatorname{pr}\left(\bigcap_{i=1}^{s} [A_{ij} = 0] \right) = \operatorname{pr}\left(\bigcap_{i=1}^{s} [\tilde{A}_{ij} = 0] \cap \bigcap_{\mu=1}^{M} [1, \dots, s \notin \mathbf{u}^{\mu} \vee j \notin \mathbf{v}^{\mu}] \right) \tag{3.6}$$

$$= (1 - \tilde{p}_1)^s (\operatorname{pr}[1, \dots, s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1])^M, \tag{3.7}$$

where we used the facts that the entries of the noise matrix $\tilde{\mathbf{A}}$ and the address patterns are generated independently of each other, and the probability that all entries of a subcolumn remain zero during learning of the μth pattern pair is independent of μ. The latter probability is written as

$$\operatorname{pr}[1, \dots, s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1] \tag{3.8}$$

$$= \operatorname{pr}([1, \dots, s \notin \mathbf{u}^1]) + \operatorname{pr}([j \notin \mathbf{v}^1]) - \operatorname{pr}([1, \dots, s \notin \mathbf{u}^1 \wedge j \notin \mathbf{v}^1]) \tag{3.9}$$

$$= \frac{\binom{m-s}{k}}{\binom{m}{k}} + \frac{\binom{n-1}{l}}{\binom{n}{l}} - \frac{\binom{m-s}{k}\binom{n-1}{l}}{\binom{m}{k}\binom{n}{l}} \tag{3.10}$$

$$= B(m,k,s) + B(n,l,1) - B(m,k,s)B(n,l,1) = 1 - \frac{l(1 - B(m,k,s))}{n}, \tag{3.11}$$

where $B(a,b,c) := \binom{a-b}{c} / \binom{a}{c} = \prod_{i=0}^{c-1} (a-b-i)/(a-i) = B(a,c,b)$; see [34] and (A.8) in the appendix. Thus

$$\operatorname{pr}(C(Y,\emptyset,j)) = \sum_{s=0}^{|Y|} (\tilde{p}_1 - 1)^s \binom{|Y|}{s} \left(1 - \frac{l}{n}(1 - B(m,k,s)) \right)^M. \tag{3.12}$$

With this result we can finally compute the general case with arbitrary, but disjunct, $Y, N = \{N_1, N_2, \ldots\} \subseteq \{1, \ldots, m\}$, $Y \cap N = \emptyset$:

$$\operatorname{pr}(C(Y,N,j)) = \operatorname{pr}(C(Y,\emptyset,j)) - \operatorname{pr}\left(\bigcup_{i=1}^{|N|} C(Y \cup \{N_i\}, \emptyset, j)\right) \tag{3.13}$$

$$= \operatorname{pr}(C(Y,\emptyset,j)) - \sum_{t=1}^{|N|} (-1)^{t+1} \sum_{1 \le i_1 < \cdots < i_t \le |N|} \operatorname{pr}\left(\bigcap_{h=1}^{t} C(Y \cup \{N_{i_h}\}, \emptyset, j)\right) \tag{3.14}$$

$$= \operatorname{pr}(C(Y,\emptyset,j)) - \sum_{t=1}^{|N|} (-1)^{t+1} \binom{|N|}{t} \operatorname{pr}(C(Y \cup \{N_1, \ldots, N_t\}, \emptyset, j)) \tag{3.15}$$

$$= \sum_{t=0}^{|N|} (-1)^t \binom{|N|}{t} \sum_{s=0}^{|Y|+t} (-1)^s \binom{|Y|+t}{s} (1-\tilde{p}_1)^s \left(1 - \frac{l}{n}(1 - B(m,k,s))\right)^M \tag{3.16}$$

$$= \sum_{s=0}^{|Y|+|N|} (1-\tilde{p}_1)^s \left(1 - \frac{l(1-B(m,k,s))}{n}\right)^M \sum_{\substack{t=\max(0,\\ s-|Y|)}}^{|N|} (-1)^{s+t} \binom{|Y|+t}{s} \binom{|N|}{t} \tag{3.17}$$

$$= \sum_{s=|N|}^{|Y|+|N|} (-1)^{s-|N|} \binom{|Y|}{s-|N|} (1-\tilde{p}_1)^s \left(1 - \frac{l}{n}(1 - B(m,k,s))\right)^M, \tag{3.18}$$

where for (3.14) we used again (A.6) (Sylvester–Poincaré), and for the last equation we used (A.7). Thus, the (Willshaw–)Palm probability for hetero-association is

$$p_{\text{Ph}}(x; k, l, m, n, M, z) = \operatorname{pr}\left(\bigcup_{Y \subseteq \tilde{u},\, |Y|=x,\, N=\tilde{u}-Y} C(Y,N,j)\right) \tag{3.19}$$

$$= \binom{z}{x} \operatorname{pr}(C(\{1,\ldots,x\}, \{x+1,\ldots,z\}, j)) \tag{3.20}$$

$$= \binom{z}{x} \sum_{s=z-x}^{z} (-1)^{s-z+x} \binom{x}{s-z+x} (1-\tilde{p}_1)^s \left(1 - \frac{l}{n}(1 - B(m,k,s))\right)^M \tag{3.21}$$

$$= \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} (1-\tilde{p}_1)^{s+z-x} \left(1 - \frac{l}{n}(1 - B(m,k,s+z-x))\right)^M \tag{3.22}$$

for $0 \le x \le z$ and B as defined below (3.11).

Now we are able to compute exact retrieval error probabilities when addressing with noisy patterns. For example, when addressing with a single address pattern containing c correct and f false one-entries and retrieving with threshold Θ, then the exact retrieval error probabilities p_{01} of a false one-entry and p_{10} of a missing one-entry are

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\text{Ph}}(x; k, l, m, n, M-1, \tilde{p}_1, c+f), \tag{3.23}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\text{Ph}}(x-c; k, l, m, n, M-1, \tilde{p}_1, f). \tag{3.24}$$

Note that the situation is as if only $M-1$ patterns were stored since, as a precondition, the pattern to be retrieved affects neither any of the synapses of a 0-neuron nor any of the synapses connecting add-noise to a 1-neuron.

3.2. Fixed pattern activity and auto-association. The analysis for hetero-association in section 3.1 can be extended to auto-association where address and content population are identical, i.e., $m=n$, $k=l$, and $u^\mu = v^\mu$ (see also appendix 1 in [34]). Here the diagonal matrix elements A_{jj} have a much higher probability,

$$\bar{p}_1 = 1-(1-\tilde{p}_1)(1-k/n)^M, \tag{3.25}$$

of being activated than nondiagonal elements (cf. (2.2)). We use again $C(Y,N,j)$ as defined in (3.1), but now we have to care whether j is contained in Y or N. We first compute the special case $N=\emptyset$ and $j \notin Y$. The analysis for $\mathrm{pr}(C(Y,\emptyset,j\notin Y))$ starts the same way as for the hetero-associative case (see (3.2)–(3.7)). Instead of (3.8)–(3.11) we have to write $\mathrm{pr}([1,\dots,s\notin u^1 \vee j\notin u^1]) = \mathrm{pr}([1,\dots,s\notin u^1]) + \mathrm{pr}([j\notin u^1]) - \mathrm{pr}([1,\dots,s,j\notin u^1]) = B(n,k,s)+B(n,k,1)-B(n,k,s+1) = 1-\frac{k}{n}(1-\frac{n}{n-s}B(n,k,s))$ and therefore

$$\mathrm{pr}(C(Y,\emptyset,j\notin Y)) = \sum_{s=0}^{|Y|}(\tilde{p}_1-1)^s\binom{|Y|}{s}\left(1-\frac{k}{n}\left(1-\frac{n}{n-s}B(n,k,s)\right)\right)^M. \tag{3.26}$$

With this result we can again compute the general case with arbitrary, but disjunct, $Y, N=\{N_1,N_2,\dots\}\subseteq\{1,\dots,m\}$, $Y\cap N=\emptyset$, but $j\notin Y\cup N$ (cf. (3.13)–(3.18)):

$$\begin{aligned}&\mathrm{pr}(C(Y,N,j\notin Y\cup N))\\ &\quad=\sum_{s=|N|}^{|Y|+|N|}(-1)^{s-|N|}\binom{|Y|}{s-|N|}(1-\tilde{p}_1)^s\left(1-\frac{k}{n}\left(1-\frac{nB(n,k,s)}{n-s}\right)\right)^M.\end{aligned} \tag{3.27}$$

If we presume $N=\emptyset$ and $j\in Y$, then (3.3) becomes $\mathrm{pr}(C(Y,\emptyset,j\in Y)) = 1-\mathrm{pr}(\bigcup_{i=1}^{|Y|-1}[A_{ij}=0]) - \mathrm{pr}[A_{jj}=0](1-\mathrm{pr}(\bigcup_{i=1}^{|Y|-1}[A_{ij}=0]|[A_{jj}=0]))$. Here the first probability on the right side evolves as before except for replacing $|Y|$ by $|Y|-1$. The conditional probability is $1-\mathrm{pr}(\bigcap_{i=1}^{|Y|-1}[A_{ij}=1]|[A_{jj}=0]) = 1-\tilde{p}_1^{|Y|-1}$ because $A_{jj}=0$ implies that the other synapses of neuron j can be activated only by noise. Thus with $\mathrm{pr}[A_{jj}=0]=1-\bar{p}_1$ we obtain

$$\mathrm{pr}(C(Y,\emptyset,j\in Y)) = \mathrm{pr}(C(Y-\{j\},\emptyset,j)) - (1-\bar{p}_1)\tilde{p}_1^{|Y|-1}. \tag{3.28}$$

This can be generalized to $N\neq\emptyset$ analogously to (3.13)–(3.18). Equation (3.15) becomes

$$\mathrm{pr}(C(Y,N,j\in Y)) = \sum_{t=0}^{|N|}(-1)^t\binom{|N|}{t}\mathrm{pr}(C(Y\cup\{N_1,\dots,N_t\},\emptyset,j\in Y)). \tag{3.29}$$

Inserting (3.28) yields two components. The first component equals (3.27) except for replacing $|Y|$ by $|Y|-1$. The second component is $\sum_{t=0}^{|N|}(-1)^t\binom{|N|}{t}(1-\bar{p}_1)\tilde{p}_1^{|Y|-1+t} = (1-\bar{p}_1)\tilde{p}_1^{|Y|-1}(1-\tilde{p}_1)^{|N|}$ and therefore

$$\begin{aligned}&\mathrm{pr}(C(Y,N,j\in Y)) = -(1-\bar{p}_1)\tilde{p}_1^{|Y|-1}(1-\tilde{p}_1)^{|N|}\\ &\quad+\sum_{s=|N|}^{|Y|+|N|-1}(-1)^{s-|N|}\binom{|Y|-1}{s-|N|}(1-\tilde{p}_1)^s\left(1-\frac{k}{n}\left(1-\frac{nB(n,k,s)}{n-s}\right)\right)^M.\end{aligned} \tag{3.30}$$

We will also need the case $j \in N$. This case implies $A_{jj} = 0$, and therefore any other synapse of neuron j can only be activated by noise. Thus simply

$$\mathrm{pr}(C(Y, N, j \in N)) = (1 - \bar{p}_1)\tilde{p}_1^{|Y|}(1 - \tilde{p}_1)^{|N|-1}. \tag{3.31}$$

With this we can finally determine the Palm probability for auto-association. If neuron j does not belong to the z address units, then we can proceed as in (3.19)–(3.22) and obtain

$$\begin{aligned} &p_{\mathrm{Pa}}(x; k, n, M, z, 0) \\ &= \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} (1 - \tilde{p}_1)^{s+z-x} \left(1 - \frac{k}{n}\left(1 - \frac{nB(n, k, s+z-x)}{n - z + x - s}\right)\right)^M . \end{aligned} \tag{3.32}$$

If neuron j is among the z address units, we have to split the union of (3.19) into two disjunct components, $\bigcup_{Y \subseteq \tilde{u},\, |Y|=x,\, N=\tilde{u}-Y,\, j \in Y} C$ and $\bigcup_{Y \subseteq \tilde{u},\, |Y|=x,\, N=\tilde{u}-Y,\, j \in N} C$. Then we can proceed again with transformations similar to (3.19)–(3.22). With (3.30), the first union corresponds to $p_{\mathrm{Pa}}(x-1; k, n, M, z-1, 0) - \binom{z-1}{x-1}(1-\bar{p}_1)\tilde{p}_1^{x-1}(1-\tilde{p}_1)^{z-x}$. With (3.31), the second union becomes $\binom{z-1}{x}(1-\bar{p}_1)\tilde{p}_1^{x}(1-\tilde{p}_1)^{z-x-1}$. Adding the two components yields

$$\begin{aligned} p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 1) &= p_{\mathrm{Pa}}(x-1; k, n, M, \tilde{p}_1, z-1, 0) \\ &\quad + (1 - \bar{p}_1)\left(p_B(x; z-1, \tilde{p}_1) - p_B(x-1; z-1, \tilde{p}_1)\right), \end{aligned} \tag{3.33}$$

and thus the general Palm probability for auto-association is

$$\begin{aligned} &p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, \sigma) \\ &= (1 - \sigma)p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 0) + \sigma p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 1). \end{aligned} \tag{3.34}$$

When addressing with a single address pattern containing c correct and f false one-entries then $\sigma = f/(n-k)$ for a lo-unit, while $\sigma = 0$ for the f noisy inputs to the hi-units. Thus, retrieving with threshold Θ, the exact retrieval error probabilities p_{01} of a false one-entry and p_{10} of a missing one-entry are

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{Pa}}(x; k, n, M-1, \tilde{p}_1, c+f, f/(n-k)), \tag{3.35}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{Pa}}(x-c; k, n, M-1, \tilde{p}_1, f, 0). \tag{3.36}$$

3.3. Random pattern activity and hetero-association. For technical applications, the patterns to be stored have often fixed pattern activities k and l (e.g., see [34, 41, 17, 42]). However, for the biological interpretation we identify the pattern activities with the size of cell assemblies [14, 5, 35], and it seems not very plausible to assume that all cell assemblies have exactly the same size. Here it might be more realistic to assume that an address pattern component is 1 with probability k/m independently of each other (and similarly l/n for the content patterns). Then the mean assembly sizes are still k and l, but the size of a given cell assemblies is a binomially distributed random variable.

The analysis can be conducted in analogy to section 3.1. Due to independently generated pattern components, (3.8) simplifies to

$$\mathrm{pr}[1,\ldots,s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1] = 1 - \frac{l}{n} + \left(\frac{l}{n}\right)\left(1-\frac{k}{m}\right)^s \tag{3.37}$$

$$= 1 - \frac{l}{n}\left(1-\left(1-\frac{k}{m}\right)^s\right). \tag{3.38}$$

Thus in the further analysis of section 3.1 we can simply replace $B(m,k,s)$ by $(1-k/m)^s$. From (3.22) we finally obtain the *Willshaw probability* for hetero-association:

$$p_{\mathrm{Wh}}(x;k,l,m,n,M,\tilde{p}_1,z)$$

$$= \binom{z}{x}\sum_{s=0}^{x}(-1)^s\binom{x}{s}(1-\tilde{p}_1)^{s+z-x}\left(1-\frac{l}{n}\left(1-\left(1-\frac{k}{m}\right)^{s+z-x}\right)\right)^M \tag{3.39}$$

$$= \sum_{i=0}^{M} p_B(i;M,l/n)\,p_B\left(x;z,1-(1-\tilde{p}_1)\left(1-\frac{k}{m}\right)^i\right) \tag{3.40}$$

for $0 \le x \le z$. The retrieval error probabilities p_{01} and p_{10} are as in (3.23), (3.24), replacing p_{Ph} by p_{Wh}. The second formula, (3.40), results from an alternative approach to obtain the Willshaw probability for random pattern activities (see [7, 6]). Here the first binomial is the probability that the considered content neuron has unit-usage i, i.e., that it has been activated i times during the learning of the M associations. Given unit usage i the term $1-(1-\tilde{p}_1)(1-k/m)^i$ is the probability that a given synapse on the content neuron has been potentiated or activated by noise. Thus, the second binomial is the probability that a content neuron receives x out of the z random inputs given a unit usage of i.

Equation (3.40) for $\tilde{p}_1 = 0$ was found in 1991 by Buckingham and Willshaw [7, 6], while (3.39) for $\tilde{p}_1 = 0$ was derived from (3.40) in 1999 by Sommer and Palm [43]. For numerical evaluations (3.39) is particularly useful if z is small and M is large, while evaluating (3.40) is more efficient for small M and large z. In cases where both M and z are large, evaluating the Willshaw probability can be computationally very expensive [22, 23].

Unfortunately, we do not know a formula for the exact Palm probability (3.22) that is analogous to (3.40). Thus, evaluating the exact error probabilities for the model variant with fixed assembly size is computationally cheap only for cases with small z. However, numerical investigations suggest that p_{W} quickly converges to p_{P} for large m, n, and z and that the resulting retrieval error probabilities for fixed assembly sizes are smaller than for random assembly size [22].

3.4. Random pattern activity and auto-association. In analogy to the previous sections we can also investigate the auto-associative case with binomially distributed pattern activities where each pattern component is active with probability k/n independently of other components. Here $\mathrm{pr}(C(Y,N,j \notin Y \cup N))$ can be obtained in the same way as done in section 3.3 for hetero-association with $k = l$ and $m = n$. This corresponds to $\sigma = 0$ and leads to $p_{\mathrm{Wa}}(x;k,n,M,\tilde{p}_1,z,0) = p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z)$. The remaining subtleties concerning autapses having a much higher activation probability $\bar{p}_1$ than other synapses (see (3.25)) can be handled in the same way as done in section 3.2 for fixed pattern activity. Thus, simply replacing $p_{\mathrm{Pa}}(x;k,n,M,\tilde{p}_1,z,0)$ by $p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z)$ we obtain from (3.34)

$$\begin{aligned} p_{\mathrm{Wa}}(x;k,n,M,\tilde{p}_1,z,\sigma) &= (1-\sigma)p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z) \\ &\quad + \sigma p_{\mathrm{Wh}}(x-1;k,n,k,n,M,\tilde{p}_1,z-1) \\ &\quad + \sigma(1-\bar{p}_1)\left(p_B(x;z-1,\tilde{p}_1) - p_B(x-1;z-1,\tilde{p}_1)\right). \end{aligned} \tag{3.41}$$

When addressing with a single address pattern containing c correct and f false one-entries then the error probabilities for threshold Θ can be computed similarly as in section 3.2,

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{Wa}}(x;k,n,M-1,\tilde{p}_1,c+f,\bar{\sigma}), \tag{3.42}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{Wa}}(x-c;k,n,M-1,\tilde{p}_1,f,0) \tag{3.43}$$

for $0 \leq x \leq z$. For the lo-units σ has to be averaged over the constrained range of possible pattern activities k' with $c \leq k' \leq n-f$; thus, $\bar{\sigma} := (\sum_{k'=c}^{n-f} p_B(k';n,k/n)f/(n-k'))/(\sum_{k'=c}^{n-f} p_B(k';n,k/n))$. Note that computing the expected Hamming distance (see (2.8)) requires a similar adjustment. Note also that p_{10} is the same as for hetero-association with the corresponding parameters (see section 3.3).

3.5. Probabilities of add-errors for pattern part retrieval. For the particular case of pattern part retrieval, $c = \lambda k$ and $f = 0$ with $0 < \lambda \leq 1$, we can use the Willshaw threshold $\Theta = \lambda k$. Then the probability of miss-errors in the retrieval outputs is generally $p_{10} = 0$. For fixed pattern activity the probability of an add-error is

$$p_{01,\mathrm{Ph}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} \left[1 - \frac{l}{n}(1 - B(m,k,s))\right]^{M-1}, \tag{3.44}$$

$$p_{01,\mathrm{Pa}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} \left[1 - \frac{k}{n}\left(1 - \frac{n}{n-s}B(n,k,s)\right)\right]^{M-1} \tag{3.45}$$

for hetero-association and auto-association, respectively. For random pattern activity, the corresponding error probabilities are

$$p_{01,\mathrm{Wh}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} \left[1 - \frac{l}{n}(1 - (1-k/m)^s)\right]^{M-1} \tag{3.46}$$

$$= \sum_{i=0}^{M-1} p_B(i;M-1,l/n)(1 - (1-\tilde{p}_1)(1-k/m)^i)^{\lambda k} \tag{3.47}$$

$$\geq [1 - (1-\tilde{p}_1)(1-kl/mn)^{M-1}]^{\lambda k} = p_1^{\lambda k}, \tag{3.48}$$

$$p_{01,\mathrm{Wa}} = p_{01,\mathrm{Wh}}|_{l=k,\, m=n}, \tag{3.49}$$

where p_B is again the binomial probability (see below (2.12)). Here the error probabilities are essentially the same for auto-association and hetero-association with $k = l$, $m = n$. Equation (3.48) corresponds to the binomial approximation equation (2.13) as used in section 2.3. The bound can be obtained from Jensen's inequality

$Ef(y) \geq f(Ey)$ (see, e.g., [9]) for convex $f(y) := (1-y)^{\lambda k}$ with random variable $y := (1-\tilde{p}_1)(1-k/m)^i$. Here the expectation $Ey = 1 - p_1$ can be computed from (A.9) using $J = 1$.

Although I could not prove this strictly, numerical experiments suggest $p_{01,\mathrm{Pa}} \leq p_{01,\mathrm{Ph}} \leq p_{01,\mathrm{Wh}} = p_{01,\mathrm{Wa}}$ [22]. The binomial approximation equation (3.48) can strongly underestimate p_{01}. Palm and Sommer [34, 38] give some asymptotic conditions when the true potential distribution converges to the corresponding binomial distribution, but only for relatively small $k \sim \log n$ and $k \leq n^{1/3}$, respectively. In section 5 we will see that the parameter range of convergence is actually much larger.

3.6. Numerical evaluations. Theorem 3.2 and the resulting retrieval error probabilities have been verified by extensive numerical simulations of the Willshaw model [22]. Some data are shown in Table 3.1.

TABLE 3.1

Results from numerical simulations of retrieval in the Willshaw model with $m = 10$, $k = 3$, $M = 5$, $\tilde{p}_1 = 0.1$ when addressing with patterns containing $c = 2$ correct and $f = 2$ false one-entries. Upper rows (S) show results for "symmetric" networks with $n = m$ and $l = k$ (cf. Figure 3.1, *left panel). Lower rows (A) show results for "asymmetric" networks with $n = 11$ and $l = 2$. The columns show optimal retrieval threshold Θ, output noise ϵ, and the error probabilities p_{01} and p_{10} for add-noise and miss-noise as well as the corresponding average values (mean) and standard errors (s.e.) from the simulation experiments (evaluating $N \approx 10^8$ retrievals in each case). The experimental values closely match the theoretical values and thus verify Theorem* 3.2.

	Θ	ϵ	p_{01}	mean	s.e.	p_{10}	mean	s.e.
S p_{Ph}	3	0.871142	0.200514	0.200473	0.000040	0.403276	0.403387	0.000049
p_{Pa}	3	0.824469	0.149855	0.149827	0.000036	0.474807	0.474726	0.000050
p_{Wh}	3	0.937330	0.223047	0.223043	0.000045	0.416887	0.416905	0.000054
p_{Wa}	4	0.974194	0.067171	0.067197	0.000027	0.817462	0.817423	0.000042
A p_{Ph}	3	1.023875	0.107831	0.107822	0.000031	0.538635	0.538590	0.000050
p_{Wh}	3	1.121372	0.127232	0.127211	0.000036	0.548828	0.548834	0.000057

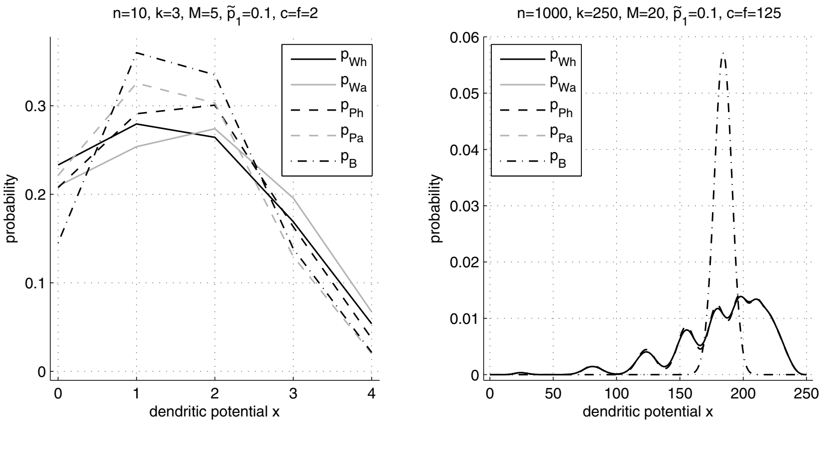


FIG. 3.1. *Examples of the Willshaw–Palm distributions (see Theorem* 3.2*) and the corresponding binomial approximation (equation* (2.11)*) for a small network (left panel) and a larger network (right panel). The plots show the distribution of the lo-units when addressing with c correct and f false units in symmetric networks ($m = n$ and $k = l$). The plots indicate that the binomial approximation can be very inaccurate.*

Figure 3.1 gives examples for the Willshaw–Palm distribution illustrating the differences between the four probability versions and the binomial approximation. For small networks the difference between the four versions of the Willshaw–Palm distribution is significant. In comparison to the binomial approximation the Willshaw–Palm probability can have a much *larger variance* and *oscillatory modulations* [19, 21]. The difference in variance is computed in section 5.2 (see (5.5)), and conditions where the variances and higher-order moments become identical are computed in section 5.4. The oscillatory modulations can be understood from (3.40) writing p_{Wh} as a superposition of $M+1$ binomials. They occur if the binomials $p_B(x;z,1-(1-\tilde{p}_1)(1-k/m)^i)$ around mean unit usage $i \approx Ml/n$ have a small standard deviation $\sqrt{z(1-\tilde{p}_1)(1-k/m)^i(1-(1-\tilde{p}_1)(1-k/m)^i)}$ compared to the mean distance $z(1-\tilde{p}_1)((1-k/m)^i-(1-k/m)^{i+1})$ between two neighboring binomials, i.e., if

$$(1-\tilde{p}_1)\frac{zk^2}{m^2}\left(1-\frac{k}{m}\right)^{Ml/n} \gg 1. \tag{3.50}$$

4. Expectation, variance, and higher-order moments of the Willshaw–Palm distribution. In this section we investigate the moments of the Willshaw–Palm probability distribution. Here we will focus on the more simple case of random pattern activity, i.e., on the Willshaw probabilities p_{Wh} and p_{Wa} (see Definition 3.1 and Theorem 3.2). The analysis for fixed pattern activity is more difficult, but it is plausible to assume that the basic (asymptotic) properties for the Palm probabilities p_{Ph}, p_{Pa} are similar to p_{Wh}, p_{Wa}. At least the expectation values of Willshaw and Palm probabilities are the same: Because the dendritic potential is $x_j = \sum_{i\in\tilde{u}} A_{ij}$, the expectation for hetero-association is identical to the corresponding binomial expectation (see section 2.3),

$$E_{p_{\mathrm{Wh}}}(x_j) = E_{p_{\mathrm{Ph}}}(x_j) = E_{p_B}(x_j) = zp_1, \tag{4.1}$$

$$E_{p_{\mathrm{Wa}}}(x_j) = E_{p_{\mathrm{Pa}}}(x_j) = zp_1 + \sigma(\bar{p}_1 - p_1), \tag{4.2}$$

where p_1 is the memory load equation (2.2). The expectation for auto-association follows similarly from $E_{p_{\mathrm{Wa}}}(x_j) = (z-1)p_1 + \sigma\bar{p}_1 + (1-\sigma)p_1$, where σ is the probability that j is among the z active units of address pattern $\tilde{u}$, and $\bar{p}_1$ is the probability that A_{jj} is active (see (3.25)).

In the following text we will sometimes write $p_0 := 1-p_1$, $\bar{p}_0 := 1-\bar{p}_1$, and $\tilde{p}_0 := 1-\tilde{p}_1$ for the sake of brevity.

4.1. Moment generating functions. The moment generating function of a random variable X with probability function p is defined by $G_p(t) := E_p(e^{tX})$ (e.g., see [39]). The following theorem shows that the moment generating functions of the Willshaw–Palm probabilities for *random* pattern activity k can be obtained from the generating function of the binomial probability (equation (A.3)).

THEOREM 4.1. *The moment generating functions* $G_{p_{\mathrm{Wh}}}(t;k,l,m,n,M,\tilde{p}_1,z)$ *and* $G_{p_{\mathrm{Wa}}}(t;k,n,M,\tilde{p}_1,z,\sigma)$ *of the Willshaw probability functions* p_{Wh} *for hetero-association (equation* (3.39)*) and* p_{Wa} *for auto-association (equation* (3.41)*) are*

$$G_{p_{\mathrm{Wh}}}(t) = \sum_{i=0}^{M} p_B(i;M,l/n)G_{p_B}(t;z,1-\tilde{p}_0(1-k/m)^i), \tag{4.3}$$

$$\begin{aligned} G_{p_{\mathrm{Wa}}}(t;\ldots,z,\sigma) &= (1-\sigma)G_{p_{\mathrm{Wh}}}(t;\ldots,z) + \sigma e^t G_{p_{\mathrm{Wh}}}(t;\ldots,z-1) \\ &\quad + \sigma\bar{p}_0(1-e^t)G_{p_B}(t;z-1,\tilde{p}_1). \end{aligned} \tag{4.4}$$

Proof. By definition it is $G_{p_{\mathrm{Wh}}}(t) := E_{p_{\mathrm{Wh}}} e^{tX} = \sum_{x=0}^{z} e^{tx} \sum_{i=0}^{M} p_B(i; M, l/n) \cdot p_B(x; z, 1-\tilde{p}_0(1-k/m)^i) = \sum_{i=0}^{M} p_B(i; M, l/n) \sum_{x=0}^{z} e^{tx} p_B(x; z, 1-\tilde{p}_0(1-k/m)^i)$. Here the second sum is the moment generating function of a binomial equation (A.3) with $N = z$ and $P = 1-(1-k/m)^i$. This shows (4.3). Similarly, the auto-associative moment generating function $G_{p_{\mathrm{Wa}}}(t)$ follows with (3.41) because moment generating functions $G_{p(x)}(t)$ are linear in $p(x)$ and have the shifting property $G_{p(x-1)}(t) = \sum_x p(x-1)e^{tx} = \sum_x p(x)e^{t(x+1)} = e^t G_{p(x)}(t)$. □

4.2. Higher order moments. The dth *raw moment* of a random variable X with probability function p is defined by the expectation $E_p X^d$ and can be computed from the moment generating function $G_p(t) := E_p(e^{tX})$, where the dth derivative $G_p^{(d)}(t)$ at $t = 0$ yields the dth moment (see, e.g., [39]). Then the dth *central moment* (or moment about the mean) is defined by the expectation $E_p(X-\mu)^d$, where $\mu := E_p X$ is the mean value. The following theorem computes the moments of the Willshaw probabilities from the moments of the binomial probability.

THEOREM 4.2. *Let $p_{\mathrm{Wh}}(x; k, l, m, n, M, \tilde{p}_1, z)$ be the Willshaw probability for hetero-association (equation* (3.39)*) and $p_B(x; z, 1-p_0)$ the corresponding binomial approximation with $p_0 := 1 - p_1$ (see* (A.2), (2.2)*). Then the raw and central moments of the Willshaw probability can be computed from the binomial moments (see* (A.4)–(A.5)*) by formally substituting powers p_0^j by numbers $p_0^{(j)}$ defined as*

$$p_0^{(j)} := \tilde{p}_0^j \left(1 - \frac{l}{n}\left(1 - \left(1 - \frac{k}{m}\right)^j\right)\right)^M, \tag{4.5}$$

where $\tilde{p}_0 := 1 - \tilde{p}_1$. For example, the raw and central Willshaw moments for hetero-association*, $\mathfrak{m}_{r,p_{\mathrm{Wh}}}(d; k, l, m, n, M, \tilde{p}_1, z)$ and $\mathfrak{m}_{c,p_{\mathrm{Wh}}}(d; k, l, m, n, M, \tilde{p}_1, z)$, can be obtained from*

$$E_{p_{\mathrm{Wh}}}(X-\mu)^d = \sum_{j=0}^{d} p_0^{(j)} (-1)^j \binom{z}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d, \tag{4.6}$$

which is true for an arbitrary offset μ. The raw and central moments follow with $\mu = 0$ and $\mu = zp_1$, respectively.

Similarly, the raw and central Willshaw moments for the auto-associative probability p_{Wa} (see (3.41)*), $\mathfrak{m}_{r,p_{\mathrm{Wa}}}(d; k, n, M, \tilde{p}_1, z, \sigma)$ and $\mathfrak{m}_{c,p_{\mathrm{Wa}}}(d; k, n, M, \tilde{p}_1, z, \sigma)$, follow from*

$$\begin{aligned} E_{p_{\mathrm{Wa}}}(X-\mu)^d &= \sum_{j=0}^{d} p_0^{(j)} (-1)^j \binom{z}{j} \left(1 - \frac{\sigma j}{z}\right) \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d \\ &+ \sigma \bar{p}_0 \sum_{j=0}^{d} \tilde{p}_0^j (-1)^j \binom{z-1}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} ((z-\mu-i-1)^d - (z-\mu-i)^d)) \end{aligned} \tag{4.7}$$

using $\mu = 0$ and $\mu = zp_1 - \sigma(\bar{p}_1 - p_1)$, respectively.

Proof. The dth *raw moment* $E_{p_{\mathrm{Wh}}} X^d$ equals the dth derivative $G_{p_{\mathrm{Wh}}}^{(d)}(t)$ at $t = 0$. From (4.3) we obtain

$$G_{p_{\mathrm{Wh}}}^{(d)}(0) = \sum_{i=0}^{M} p_B(i; M, l/n) G_{p_B}^{(d)}(0; z, 1-\tilde{p}_0(1-k/m)^i),$$

where $G^{(d)}_{p_B}(0;N,1-Q) = \mathfrak{m}_{r,p_B}(d,N,1-Q) = \sum_{j=0}^{d} c_j^{(d)}(N)Q^j$ is the dth raw moment of the binomial probability (see (A.5)). For brevity we have defined coefficients $c_j^{(d)}(N) := (-1)^j\binom{N}{j}\sum_{k=0}^{j}(-1)^k\binom{j}{k}(N-k)^d$. Applying (A.9) we obtain

$$\begin{aligned} E_{p_{\mathrm{Wa}}}X^d &= \sum_{i=0}^{M} p_B(i;M,l/n)\sum_{j=0}^{d} c_j^{(d)}(z)\tilde{p}_0^j(1-k/m)^{ij} \\ &= \sum_{j=0}^{d} c_j^{(d)}(z)\tilde{p}_0^j\sum_{i=0}^{M} p_B(i;M,l/n)(1-k/m)^{ij} = \sum_{j=0}^{d} c_j^{(d)}(z)p_0^{(j)}. \end{aligned}$$

This proves the formulae for the raw moments $\mathfrak{m}_{r,p_{\mathrm{Wh}}}$, for example, (4.6) for $\mu = 0$. The general moment equation (4.6) with arbitrary offset μ follows then from inserting the raw moments into $E(X-\mu)^d = \sum_{i=0}^{d}\binom{d}{i}(-\mu)^{d-i}EX^i$, where we used the binomial sum (see below (A.9)) and the linearity of the expectation operator. Inserting $\mu = zp_1$ (see (4.1)) finally yields the central moments $\mathfrak{m}_{c,p_{\mathrm{Wh}}}$ for the hetero-associative Willshaw probability (see also (A.5); cf. [25]).

Similarly, the general moment equation (4.7) for auto-association follows with (3.41) because moments $E_{p(x)}(X-\mu)^d$ are linear in $p(x)$ and have the shifting property $E_{p(x-1)}(X-\mu)^d = \sum_x p(x-1)(x-\mu)^d = \sum_x p(x)(x-\mu+1)^d = E_{p(x)}(X-(\mu-1))^d$. In particular, summing the two Willshaw terms in (3.41) leads to the factor $(1-\sigma)\binom{z}{j} + \sigma\binom{z-1}{j} = \binom{z}{j}(1-\frac{\sigma j}{z})$ in (4.7). The raw and central moments $\mathfrak{m}_{r,p_{\mathrm{Wa}}}$ and $\mathfrak{m}_{c,p_{\mathrm{Wa}}}$ then follow from inserting $\mu = 0$ and $\mu = E_{p_{\mathrm{Wa}}}X$ (see (4.2)). □

The following lemma gives a more detailed characterization of the numbers $p_0^{(j)}$ that have been used to compute the moments of the Willshaw probability.

Lemma 4.3. *Let $p_0^{(j)}$ be as defined in Theorem 4.2. For $0 < P < 1$ we have*

$$R_j(P) := \frac{1}{P}\sum_{i=2}^{j}\binom{j}{i}(-P)^i = \frac{(1-P)^j - 1 + Pj}{P} \geq 0, \tag{4.8}$$

$$p_0 := 1 - p_1 = \tilde{p}_0\left(1 - \frac{kl}{mn}\right)^M, \tag{4.9}$$

$$p_0^{(j)} := \tilde{p}_0^j\left(1 - \frac{l}{n}\left(1 - \left(1 - \frac{k}{m}\right)^j\right)\right)^M = \tilde{p}_0^j\left(1 - \frac{kl}{mn}\left(j - R_j\left(\frac{k}{m}\right)\right)\right)^M \approx p_0^j. \tag{4.10}$$

For $j = 0, 1$ we have $R_j(P) = 0$ and $p_0^{(j)} = p_0^j$. For $j \geq 2$ we have $R_j(P) > 0$. For sufficiently small $P \to 0$ the bound $R_j(P) < \binom{j}{2}P$ becomes true. Furthermore, for $j \geq 2$ we have the bounds

$$p_0^j < p_0^{(j)} < p_0^{j-R_j(k/m)}, \tag{4.11}$$

$$0 < \frac{p_0^{(j)} - p_0^j}{p_0^j} < p_0^{-R_j(k/m)} - 1 < -(e-1)R_j(k/m)\ln p_0, \tag{4.12}$$

where the latter bound in (4.12) *is true only for $-R_j(k/m)\ln p_0 < 1$. In particular, the relative difference between $p_0^{(j)}$ and p_0^j vanishes for $R_j(k/m)\ln p_0 \to 0$. Finally, let $p := k/m \to 0$, $q := l/n$, $M = \ln p_0/\ln(1-pq)$ (see* (2.3)*). Then for $j^2p(1-\ln p_0) \to 0$,* *fixed* $\tilde{p}_1$*, and using the asymptotic* Θ *notation as defined in the appendix, we have*

$$p_0^{(j)} - p_0^j = -\binom{j}{2} p(1-q) p_0^j \ln p_0 + \Theta(j^3 p^2 p_0^j (1 - j \ln p_0) \ln p_0). \tag{4.13}$$

Proof. Equation (4.8) follows from the binomial sum (see below (A.9)). Equation (4.9) is simply rewriting (2.2) with $\tilde{p}_0 := 1 - \tilde{p}_1$ for the sake of completeness. Equation (4.10) follows from simple transformations of the definitions (4.5), (4.8). The claims for $j = 0, 1$ follow trivially. $R_j(P) > 0$ for $j \geq 2$ follows from $(1-P)^j > (1 - Pj)$ (see (A.11)). $R_j(P) < \binom{j}{2} P$ for $P \to 0$ follows directly from the definition of R_j. The lower bound in (4.11) follows from $(1 - kl/mn)^j = 1 - (kl/mn)(j - R_j(kl/mn))$ because $R_j(P)$ is monotonically increasing for $0 < P < 1$. The upper bound in (4.11) follows from $(1 - pq(j - R_j)) < (1-pq)^{j - R_j}$ (see (A.11) with $j - R_j > j - R_j(1) = 1$). Equation (4.12) follows from (4.11) and (A.12). We finally prove the asymptotic approximation equation (4.13): For $p \to 0$, $M = \ln p_0 / \ln(1 - pq)$ (see (2.3)) we have with (A.13)–(A.14)

$$M = \frac{-\ln p_0}{pq + \Theta(p^2 q^2)} = \frac{-\ln p_0}{pq} \frac{1}{1 + \Theta(pq)} = \frac{-\ln p_0}{pq}(1 + \Theta(pq)), \tag{4.14}$$

$$Mpq = -\ln p_0 + \Theta(pq \ln p_0), \tag{4.15}$$

$$R_j(p) = \binom{j}{2} p - \binom{j}{3} p^2 + \cdots = \binom{j}{2} p + \Theta(j^3 p^2) \to 0 \text{ for } j^2 p \to 0. \tag{4.16}$$

The final purpose of this is to find a close approximation for

$$p_0^{(j)} - p_0^j = p_0^j \left(\frac{p_0^{(j)}}{p_0^j} - 1 \right) \quad \text{with} \quad \frac{p_0^{(j)}}{p_0^j} = e^{M(\ln(1 - pq(j - R_j)) - j \ln(1 - pq))}. \tag{4.17}$$

For $R_j \to 0$ the term in the outer brackets of the exponential is written as $\ln(1 - pq(j - R_j)) - j \ln(1 - pq) = -pq(j - R_j) - \frac{p^2 q^2 (j - R_j)^2}{2} + \Theta(p^3 q^3 j^3) - j(-pq - \frac{p^2 q^2}{2} + \Theta(p^3 q^3)) = pqR_j - \frac{p^2 q^2}{2} \left((j - R_j)^2 - j \right) + \Theta(p^3 q^3 j^3)$. Here we have $(j - R_j)^2 - j = j^2 - j + \Theta(j R_j) = j^2 - j + \Theta(j^3 p)$ and therefore

$$\frac{p_0^{(j)}}{p_0^j} = e^{M(pqR_j - 0.5 p^2 q^2 (j^2 - j) + \Theta(p^3 q^2 j^3))} \tag{4.18}$$

$$= e^{M(p^2 q \binom{j}{2}(1-q) + \Theta(p^3 q j^3))} = e^{Mpq(p\binom{j}{2}(1-q) + \Theta(p^2 j^3))} \tag{4.19}$$

$$= e^{(-\ln p_0 + \Theta(pq \ln p_0))(p\binom{j}{2}(1-q) + \Theta(p^2 j^3))} = e^{-\binom{j}{2} p(1-q) \ln p_0 + \Theta(j^3 p^2 \ln p_0)} \tag{4.20}$$

$$= 1 - \binom{j}{2} p(1-q) \ln p_0 + \Theta(j^4 p^2 \ln^2 p_0 - j^3 p^2 \ln p_0) \tag{4.21}$$

$$= 1 - \binom{j}{2} p(1-q) \ln p_0 + \Theta(j^3 p^2 \ln p_0 (1 - j \ln p_0)). \qquad \square \tag{4.22}$$

4.3. Variance. Applying Theorem 4.2, we can easily compute the second raw and central moments of the Willshaw probability from the well-known second moments of the corresponding binomial probability $p_B(x; z, 1 - p_0)$ (see also (A.4)–(A.5)). Thus, replacing p_0^j by $p_0^{(j)}$ in

$$E_{p_{B(x;z,1-p_0)}} X^2 = z p_0 (1 - p_0) + z^2 (1 - p_0)^2 = z^2 + p_0 (z - 2z^2) + p_0^2 (z^2 - z) \tag{4.23}$$

gives us immediately the second moment and variance of the Willshaw probability p_{Wh} for *hetero-association*,

$$\begin{align} E_{p_{\mathrm{Wh}}} X^2 &= z^2 + p_0(z - 2z^2) + p_0^{(2)}(z^2 - z) \tag{4.24}\\ &= z^2(1 - 2p_0 + p_0^{(2)}) + z(p_0 - p_0^{(2)}), \tag{4.25}\\ \mathrm{Var}_{p_{\mathrm{Wh}}} X = E_{p_{\mathrm{Wh}}} X^2 - E^2_{p_{\mathrm{Wh}}} X &= z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)}), \tag{4.26}\end{align}$$

where $p_0^{(2)} = \tilde{p}_0^2(1-(kl/mn)(2-k/m))^M$. Note that in (4.25), (4.26) all coefficients of z are positive since with (4.11) we have $p_0 > p_0^{2-k/m} > p_0^{(2)} > p_0^2$ and $1 - 2p_0 + p_0^{(2)} > (1 - p_0)^2 > 0$ for $0 < k/m, l/n < 1$. Thus, the variance increases monotonically with z. Indeed, the variance of the Willshaw probability scales with z^2, while the corresponding binomial variance scales only with z [21]. Applying (4.7) with $d = 2$, we easily obtain the second moments of the Willshaw probability p_{Wa} for *auto-association*,

$$\begin{align} E_{p_{\mathrm{Wa}}}(X - \mu)^2 = (z - \mu)^2 &- (2(z - \mu) - 1)(zp_0 - \sigma(p_0 - \bar{p}_0)) \nonumber\\ &+ (z - 1)((z - 2\sigma)p_0^{(2)} + 2\sigma\bar{p}_0\tilde{p}_0), \tag{4.27}\end{align}$$

$$\begin{align} \mathrm{Var}_{p_{\mathrm{Wa}}} X &= z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)} - 2\sigma(p_0^{(2)} - p_0^2 - \bar{p}_0(\tilde{p}_0 - p_0))) \nonumber\\ &\qquad - \sigma(p_0 - 2p_0^{(2)} + \sigma(p_0 - \bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)), \tag{4.28}\\ &= \mathrm{Var}_{p_{\mathrm{Wh}}} X - 2\sigma z(p_0^{(2)} - p_0^2 - \bar{p}_0(\tilde{p}_0 - p_0)) \nonumber\\ &\qquad - \sigma(p_0 - 2p_0^{(2)} + \sigma(p_0 - \bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)). \tag{4.29}\end{align}$$

Here (4.27) is true for any offset μ. In particular, the second raw moment follows with $\mu = 0$, and the variance equation (4.28) follows with $\mu = E_{p_{\mathrm{Wa}}} X = z - zp_0 + \sigma(p_0 - \bar{p}_0)$ (see (4.2)).

4.4. Auto-association vs. hetero-association. As long as the dendritic potential distribution has a Gaussian shape, the variance determines retrieval quality, i.e., the larger the variance, the larger the error probabilities (e.g., see (3.42)). Thus, in order to answer the question whether retrieval quality is better for auto-association or hetero-association (with $k = l$ and $m = n$), the following lemma investigates the asymptotic behavior of the variances difference $\delta_{\mathrm{Var_{WaWh}}} := \mathrm{Var}_{p_{\mathrm{Wa}}} X - \mathrm{Var}_{p_{\mathrm{Wh}}} X$. To obtain general results, we fix the memory load $p_1 = p_{1\epsilon}$ to its maximum under quality constraint ϵ, as discussed in section 2.3 (see (2.15)).

Lemma 4.4. *For $p = k/n \to 0$, $\sigma/p \sim 1$, $z \sim k$, $p \ln p \ln k \to 0$, $z \sim k$, fixed $\tilde{p}_1$, and "hifi" memory load $1 - p_0 = p_{1\epsilon} = (\epsilon p)^{1/z}$ as in* (2.15), *we have*

$$\begin{align} \delta_{\mathrm{Var_{WaWh}}} &\approx 2\sigma z p p_0^2 \ln p_0 - \sigma(p_0 - 2p_0^2) \tag{4.30}\\ &\approx \begin{cases} +\sigma, & p_0 \to 1, \\ -\sigma p_0, & p_0 \to 0. \end{cases} \tag{4.31}\end{align}$$

Proof. We can apply (4.13) because $p \ln p \ln k \to 0$ implies $p \ln p_0 \to 0$ even for $p_0 \to 0$ with $p_0 \approx -\ln(\epsilon p)/z$ (see (2.20)). Thus, (4.30) follows from (4.29) because $p_0 - 2p_0^{(2)} \sim -pp_0^2 \ln p_0$ dominates over $\bar{p}_0 \sim e^{-Mp} = e^{(\ln p_0)/p} = p_0^{n/k}$ even for sparse potentiation with $p_0 \to 1$ and $\bar{p}_0 \sim (1 - (\epsilon k/n)^{1/z})^{n/k} \sim \exp(-(\epsilon k/n)^{1/z}(n/k))$, and similarly, $p_0 - 2p_0^{(2)} \approx p_0 - 2p_0^2$ dominates over $\sigma(p_0 - \bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)$.

Equation (4.31) follows because for sparse potentiation with $p_0 \to 1$ we have $p_0^2 \ln p_0 \to 0$ and $zp \sim k^2/n \to 0$ (see (2.19)). Similarly, for dense potentiation with

$p_0 \to 0$ we have $-\sigma(p_0 - 2p_0^2) \approx -\sigma p_0$, and for $p_0 \approx -\ln(\epsilon p)/z$ (see (2.20)) we have $0 > 2\sigma z p p_0^2 \ln p_0 \approx 2\sigma p_0(-p\ln(\epsilon p)\ln z)$. $\square$

Thus, the auto-associative variance $\mathrm{Var}_{p_{\mathrm{Wa}}}X$ becomes larger than $\mathrm{Var}_{p_{\mathrm{Wh}}}X$ for sparse potentiation with $p_1 \to 0$, but smaller for dense potentiation with $p_1 \to 1$. However, remember from sections 3.4 and 3.2 that pattern part retrieval with $f = 0$ (i.e., no add-errors in the address pattern) implies $\sigma = 0$ and thus identical distributions for auto-association and hetero-association. Also, the following lemma shows that in general the differences between hetero-associative and auto-associative moments vanish asymptotically.

LEMMA 4.5. *For fixed d, $p = k/n \to 0$, $\sigma/p \sim 1$, $z \sim k$, $zp_1 \leq \mu \leq z$, $p\ln p_0 \to 0$, and "hifi" memory load $1 - p_0 = p_{1\epsilon} = (\epsilon p)^{1/z}$ as in* (2.15), *we have*

$$E_{p_{\mathrm{Wh}}}(X-\mu)^d - E_{p_{\mathrm{Wa}}}(X-\mu)^d \to 0. \tag{4.32}$$

Proof. The difference is identical to (4.7) (cf. (4.6)) except that the factor $(1-\frac{\sigma j}{z})$ in the first double sum becomes simply $\frac{\sigma j}{z} \sim \frac{j}{n}$. Thus, with (A.10) the absolute value of the first double sum becomes zero because

$$\left| \sum_{j=0}^{d} p_0^{(j)} (-1)^j \frac{\sigma j}{z} z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} b_{di} (z-\mu-j)^{\underline{i-j}} \right|$$
$$\leq \sum_{j=0}^{d} p_0^{(j)} \frac{\sigma j}{z} z^j \sum_{i=j}^{d} \binom{i}{j} S_{di} (zp_0)^{i-j} \sim \frac{(zp_0)^d}{n} \to 0.$$

The inequality is true for large enough z with $d < z-\mu \leq z - zp_1 = zp_0$. The asymptotic approximations remain true even for small constant z where we used $p_0^{(j)} \sim p_0^j$ (see (4.12)). Note here that $z = O(\log n)$ implies sparse or balanced potentiation with $1 \geq p_0 \not\to 0$, while larger z implies dense potentiation with $p_0 \to 0$ and $zp_0 = O(\log n)$ (see (2.19), (2.20)). We still have to show that also the second double sum in (4.7) becomes zero:

$$\left| \sigma \bar{p}_0 \sum_{j=0}^{d} \tilde{p}_0^j (-1)^j \binom{z-1}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} ((z-\mu-i-1)^d - (z-\mu-i)^d) \right|$$
$$\sim O\left(\frac{\bar{p}_0 z^{2d+1}}{n}\right) \to 0.$$

This is obvious for sparse and balanced potentiation when $z \sim k \sim O(\log n)$ (see (2.19)), but follows also for dense potentiation ($p_0 = O((\log n)/z) \to 0$; see (2.20)) since here $\bar{p}_0 \sim e^{-Mp} = e^{(\ln p_0)/p} = p_0^{n/k}$ quickly approaches zero. $\square$

5. Comparison of Willshaw–Palm to binomial distribution. In this section we compare the Willshaw–Palm probability distribution of the dendritic potentials (see Definition 3.1) to the corresponding binomial approximation $p_B(x; z, p_1)$, which assumes independently generated memory matrix entries (see section 2.3). In particular, we are interested in asymptotic conditions when the two probability distributions, as judged by their moments, become identical for maximal memory load (i.e., $p_1 = p_{1\epsilon}$ and $M = M_\epsilon$ as estimated by (2.15), (2.16)). This corresponds to correctness conditions for many previous results that rely on the binomial approximation equation (2.13) (see, e.g., [46, 34, 33, 37, 7, 38, 4, 43, 20]).

5.1. Difference in moments. For the difference $\Delta^{(d)}_{\mathrm{Wh}}$ between the dth moments of the *hetero-associative* Willshaw probability p_{Wh} and the corresponding binomial probability $p_B(x; z, 1-p_0)$ (see section 2.3), we obtain from (4.6), (A.5), (A.10)

$$\Delta^{(d)}_{\mathrm{Wh}}(\mu) := E_{p_{\mathrm{Wh}}}(X-\mu)^d - E_{p_B(x;z,1-p_0)}(X-\mu)^d \tag{5.1}$$

$$= \sum_{j=2}^{d} (p_0^{(j)} - p_0^j)(-1)^j \binom{z}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d \tag{5.2}$$

$$= \sum_{j=2}^{d} (p_0^{(j)} - p_0^j)(-1)^j z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di} (z-\mu-j)^{\underline{i-j}}, \tag{5.3}$$

where μ is again an arbitrary offset (e.g., $\mu = 0$ for the raw moments and $\mu = zp_1$ for the central moments). Thus, $\Delta^{(d)}_{\mathrm{Wh}}$ has the same form as the dth binomial moment, written as a polynomial in p_0, but where powers p_0^j have been replaced by $p_0^{(j)} - p_0^j$ (see Theorem 4.2). Also note that $p_0^{(j)} = p_0^j$ for $j = 0, 1$ (see Lemma 4.3). The corresponding difference $\Delta^{(d)}_{\mathrm{Wa}}(\mu) := E_{p_{\mathrm{Wa}}}(X-\mu)^d - E_{p_{\mathrm{B}}(x;z,1-p_0)}(X-\mu)^d$ for the *auto-associative* Willshaw probability p_{Wa} can be obtained in a similar way from (4.7).

5.2. Difference in variance. A particularly interesting case is variance ($d = 2$): As long as the overall distribution of dendritic potentials resembles a Gaussian (which is often true, but see [19, 21, 23]), the retrieval error probabilities are essentially determined by the first two moments, i.e., expectation ($d = 1$) and variance ($d = 2$). Thus, it seems plausible to assume that a necessary condition for convergence of the Willshaw–Palm distribution towards a binomial is that expectation and variance become identical. For hetero-association, the expectations are already identical (equation (4.1)). Thus, it is sufficient to investigate conditions when the difference $\delta_{\mathrm{Wh}B}$ between the two variances vanishes. From (4.26), (4.23) we obtain for hetero-association

$$\delta_{\mathrm{Wh}B} := \mathrm{Var}_{p_{\mathrm{Wh}}} X - \mathrm{Var}_{p_B} X = z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)}) - zp_0(1-p_0) \tag{5.4}$$

$$= (z^2 - z)(p_0^{(2)} - p_0^2) > 0. \tag{5.5}$$

Note that $\delta_{\mathrm{Wh}B}$ is always positive (see (4.11)). Thus, the binomial approximation always underestimates the variance of the dendritic potentials. Therefore the binomial approximation generally underestimates the probabilities of retrieval errors and overestimates storage capacity, at least if the Willshaw distribution comes close to a Gaussian, which is often true (cf. (3.48) for pattern part retrieval; but see [19, 21, 23]). With (4.12), (4.13) we obtain

$$\delta_{\mathrm{Wh}B} \le (z^2 - z)p_0^2(p_0^{-k/m} - 1) \approx -(z^2 - z)\frac{k}{m}\left(1 - \frac{l}{n}\right) p_0^2 \ln p_0, \tag{5.6}$$

where the approximation is true for $(k/m)(1 - \ln p_0) \to 0$ (see also (2.20)). Note that (5.6) can become zero for a very large parameter range under maximal memory load (see (2.19), (2.20)).

The analysis for auto-association is similar. The difference $\delta_{\mathrm{Wa}B} := \mathrm{Var}_{p_{\mathrm{Wa}}}(X) - \mathrm{Var}_{p_B}(X)$ can be obtained from (5.5) and (4.29). It is easy to see from (4.32) that in general $\delta_{\mathrm{Wa}B}$ vanishes asymptotically with $\delta_{\mathrm{Wh}B}$. In the following two sections we generalize our asymptotic considerations to higher-order moments.

5.3. Convergence of the raw moments. The following lemma determines asymptotic conditions when the dth raw moment of the Willshaw–Palm probability p_{Wh} becomes identical to the dth raw moment of the corresponding binomial probability $p_B(x;z,1-p_0)$, i.e., conditions when the difference $\Delta^{(d)}_{\mathrm{Wh}}(0) := E_{p_{\mathrm{Wh}}}X^d - E_{p_B(x;z,1-p_0)}X^d$ becomes zero.

LEMMA 5.1. *For fixed d and $(k/m)\ln p_0 \to 0$ the following bounds become asymptotically true:*

$$|\Delta^{(d)}_{\mathrm{Wh}}(0)| \le \sum_{j=2}^{d} (p_0^{(j)} - p_0^j) z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di} (z-j)^{\underline{i-j}} \tag{5.7}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} z^{\underline{j}} p_0^j \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(z-j)^{\underline{i-j}} \tag{5.8}$$

$$\le -d^2\frac{k}{m} z^d p_0^2 \ln p_0 \sum_{j=2}^{d}\sum_{i=j}^{d}\binom{i}{j} S_{di} \sim \frac{k}{m} z^d p_0^2 \ln p_0 \le \frac{kz^d}{m}. \tag{5.9}$$

Proof. The lemma follows from (5.3), (4.12), and $R_j(k/m) < \binom{j}{2}(k/m)$ (see Lemma 4.3). □

Thus, the raw moments of the Willshaw–Palm probability p_{Wh} and the corresponding binomial probability $p_B(x;z,1-p_0)$ become identical if the address pattern activities $k := |\mathbf{u}^\mu|$ and $z := |\tilde{\mathbf{u}}|$ grow at most polynomial in the logarithm $\log m$ of the address population size m.

In the following section we will see that even for larger $k(m)$ the two probability distributions can still become essentially identical as judged by the difference of the *central* moments. The reason for this effect can be easily explained: Consider two probability distributions p_A and p_B with zero mean values and $\delta(x) := p_A(x) - p_B(x) \to 0$ and also $\delta(x)/p_A(x) \to 0$ and $\delta(x)/p_A(x) \to 0$ for any x. Then assume that the dth (central) moments converge, i.e., $\sum x^d \epsilon(x) \to 0$. Then it is still possible that the corresponding distributions p'_A and p'_B with mean $\mu > 0$ have diverging moments because $\sum (x+\mu)^d \epsilon(x)$ can grow arbitrarily with μ. This is the motivation to have a closer look at the convergence of the central moments in the following section.

5.4. Convergence of the central moments. Here we determine asymptotic conditions when the dth *central* moments of the Willshaw–Palm probabilities p_{Wh} and p_{Wa} become identical to the dth central moment of the corresponding binomial probability $p_B(x;z,1-p_0)$, i.e., conditions when $\Delta^{(d)}_{\mathrm{Wh}}(\mu)$ and $\Delta^{(d)}_{\mathrm{Wa}}(\mu)$ become zero (see section 5.1).

LEMMA 5.2. *For fixed d, $(k/m)\ln p_0 \to 0$, and $d < z - \mu \le z - zp_1 = zp_0$ the following bounds become asymptotically true:*

$$|\Delta^{(d)}_{\mathrm{Wh}}(\mu)| \le \sum_{j=2}^{d} (p_0^{(j)} - p_0^j) z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di} (z-\mu-j)^{\underline{i-j}} \tag{5.10}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} z^{\underline{j}} p_0^j \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(zp_0)^{\underline{i-j}} \tag{5.11}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(zp_0)^{i} \tag{5.12}$$

$$(5.13) \qquad \leq -d^2 \frac{k}{m}(zp_0)^d \ln p_0 \sum_{j=2}^{d} \sum_{i=j}^{d} \binom{i}{j} S_{di} \sim \frac{-k(zp_0)^d \ln p_0}{m}.$$

Proof. The lemma follows from (5.3), (4.12), and $R_j(k/m) < \binom{j}{2}(k/m)$ (see Lemma 4.3). □

With this we can easily find asymptotic convergence conditions for maximal memory load as approximately analyzed in section 2.3.

THEOREM 5.3. *For maximal memory load as estimated by the binomial approximation* (2.13)*, i.e., for $p_1 = p_{1\epsilon}$ and $M = M_\epsilon$ as estimated by* (2.15), (2.16)*, the dth central moment of the Willshaw–Palm probability p_{Wh} becomes identical to the dth central moment of the corresponding binomial probability $p_B(x; z, p_{1\epsilon})$, i.e.,*

$$(5.14) \qquad \Delta_{\mathrm{Wh}}^{(d)}(zp_{1\epsilon}) \to 0 \quad \textit{if} \quad \frac{k(\ln \frac{n}{\epsilon l})^d \ln z}{m} \to 0.$$

Thus, for n polynomial in m the dth central moments converge at least for $k = O(m/\log^{d+2} m)$. In particular, the variances converge at least for $k = O(m/\log^4 m)$. Moreover, all *central moments converge, and therefore the Willshaw probability p_{Wh} becomes identical to the binomial approximation, at least for $k = O(m^P)$ with fixed $P < 1$.*

Proof. For $zp_{0\epsilon} \to \infty$ with $p_{0\epsilon} := 1 - p_{1\epsilon}$, the theorem follows from Lemma 5.2 by using $zp_{0\epsilon} \sim \log(n/(\epsilon l))$ (see (2.20)). For smaller (e.g., constant) z the convergence of the central moments follows already from the convergence of the raw moments (see Lemma 5.1). □

A particular case are "symmetric" networks with $m = n$ and $k = l$, for example, for auto-association. It turns out that for such networks the range of convergence is even larger: Assume $k = n/\ln^P n$. Then the convergence condition in (5.14) becomes $(k/n)(\ln((\ln^P n)/\epsilon))^d \ln n \to 0$. Thus, here the central moments converge at least for $k = O(n/\log^2 n)$. Note that the results for hetero-association apply also to auto-association due to Lemma 4.5. Together these considerations suggest that the theoretical results on neural associative networks apply to a much larger range than assumed previously [34, 38, 19]. This includes large portions of the dense potentiation regime with $k/\log n \to \infty$ (see section 2.4). Here previous analyses relying on the binomial approximation have suggested the potential for very efficient computer implementations and new biological hypotheses about the roles of structural plasticity and inhibitory neurons [22, 24].

5.5. Numerical evaluations. The results of this section are verified by Figure 5.1, which shows data from numerical experiments testing how well the binomial theory approximates exact values. In fact, the reliability of the binomial approximation depends both on the network size (n) and the pattern activity (k). In general, the binomial theory becomes better for larger n and smaller k. The approximations of pattern capacity M_ϵ (equation (2.16)) and network capacity C_ϵ (equation (2.17)) are comparably reliable and, even for linear $k = n/2$ and small n, overestimate the true values by less than a factor of two (Figure 5.1a).

However, the derived "compression capacities" C^I and C^S depend on the maximal memory load $p_{1\epsilon}$ (or $1 - p_{1\epsilon}$; see sections 2.2 and 2.4), which can be strongly overestimated by the binomial theory (Figure 5.1b). For linear $k \sim n$ the relative error seems to grow without bound, implying $C^I \to 0$ and possibly $C^S \to 0$. Nevertheless, for smaller k the binomial approximation is much better already for realistic

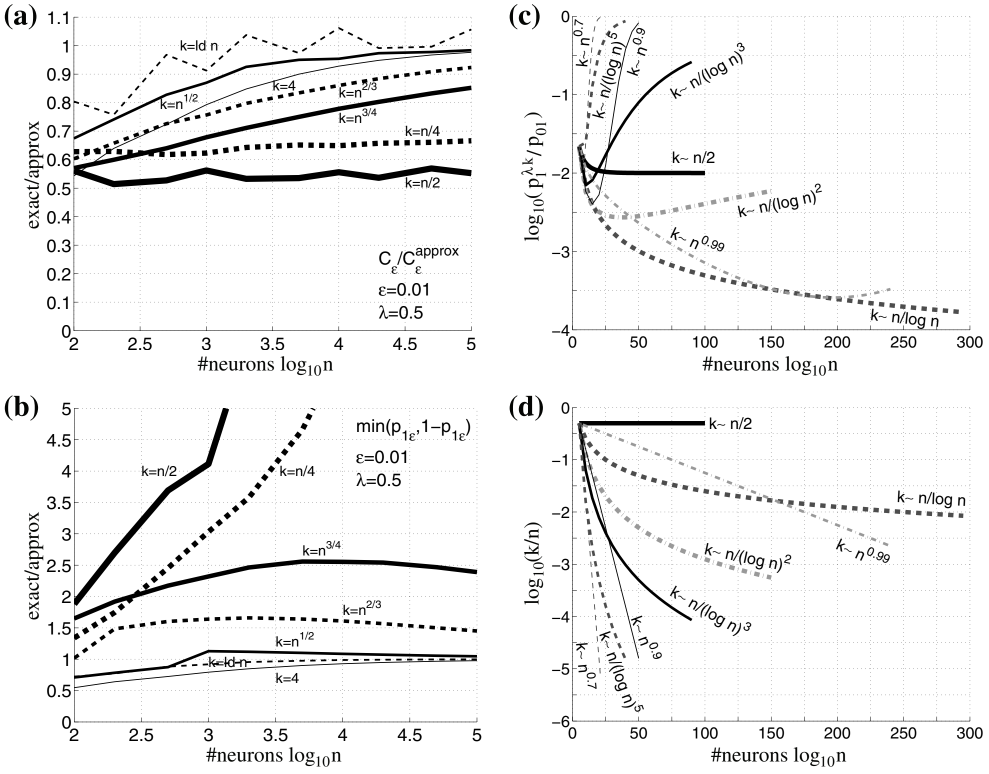


FIG. 5.1. *Numerical experiments comparing the binomial approximative analysis to the exact theory for $m = n$, $k = l$, $\tilde{p}_1 = 0$, $\epsilon = 0.01$, and pattern part retrieval with half addresses ($\lambda = 0.5$).* (a) *Relative approximation error for network storage capacity, $C_\epsilon/C_\epsilon^{\text{approx}}$ (see* (2.10), (2.17)*). Each curve corresponds to a particular pattern activity function $k(n)$ growing with the neuron number n (log-scale) as indicated in the plots. Relative errors for pattern capacity M_ϵ are virtually identical.* (b) *Relative approximation error for the memory load $p_{1\epsilon}$ at maximal pattern load M_ϵ (see* (2.9), (2.2)*), similar to panel* (a)*. More exactly, the plots show* $\min(p_{1\epsilon}, 1-p_{1\epsilon})/\min(p_{1\epsilon}^{\text{approx}}, 1-p_{1\epsilon}^{\text{approx}})$, *where $p_{1\epsilon}^{\text{approx}}$ is the approximation* (2.15)*. The corresponding approximation errors for the related compression capacities C_ϵ^I and C_ϵ^S (see section* 2.4*) look qualitatively very similar (cf.* [23]*).* (c) *Relative approximation error (log-scale) for the retrieval error probability* p_{01} *(see* (2.13), (3.47)*) when storing M_ϵ patterns approximated by* (2.16)*. Each curve corresponds to a particular function $k(n)$ with $k(10^5) = 50000$. Each case was evaluated for increasing n until a maximal computation time was reached (about* 50*h per data point on a* 2.4*GHz AMD Opteron processor evaluating relevant summands of* (3.47) *with computing precision* 1000*bit (see* [23] *for further details). The plots indicate convergence $p_1^{\lambda k}/p_{01} \to 1$ for $k = O(n/\log^2 n)$, but divergence for $k \sim n/\log n$, thus verifying the theoretical results of section* 5.4. (d) *Actual pattern activities k/n (log-scale) corresponding to panel* (c).

network sizes. For example, for $n = 10^5$ the information capacity C^I is about 100% of the binomial estimate for constant $k = 4$, 95% for $k = n^{1/2}$, 70% for $k = n^{2/3}$, and still 40% for $k = n^{3/4}$ (similar values for C^S; data not shown). Interestingly, the binomial approximations first become worse with growing n until a turning point is reached (e.g., $n = 10^4$ for $k = n^{3/4}$), and only then approach finally the exact values.

Figure 5.1c, d shows results for very large network size n and comparably large pattern activities $k(n)$. For near linear $k(n)$ the turning points are reached only for n too large to be useful for applications or relevant for biology. Nevertheless, for smaller pattern activities, for example, $k = O(n^{0.8})$, the convergence is much faster. Turning points as described above are still visible for $k = O(n/\log^2 n)$, but seem absent for

$k \sim n/\log n$. Thus, the numerical experiments are consistent with the theoretical bound derived at the end of section 5.4.

6. Conclusions.

Theories on neural associative networks with binary synapses often use a binomial approximation of the dendritic potential distribution to estimate retrieval error probabilities and performance measures such as storage capacity or retrieval speed [46, 34, 37, 33, 4, 43, 20]. However, for finite network size n or patterns with a relatively large number of active units k this approximation can be very inaccurate. So far, the convergence of the binomial approximation to the true potential distribution and thus the asymptotic correctness of the classical theory has been demonstrated only for some special cases involving very sparse activity patterns, where a binary pattern vector of n neurons contains on average only $k = \log n$ or $k \leq n^{1/3}$ active units [34, 38]. This appeared sufficient because it was believed that neural associative networks would be efficient only for extreme sparseness anyway [34, 43]. In contrast, recent applications of the theory to problems requiring less sparse patterns has gained increased attention for a number of reasons described in the introduction. For example, theoretical analyses based on the binomial approximation suggest that associative networks can operate very efficiently for large pattern activities with $k/\log n \to \infty$ (or equivalently "dense potentiation" with memory load $p_1 \to 1$) if the synaptic matrix is adequately compressed [18, 19, 20, 22]. However, the correctness of these results has been doubted because it remained unclear whether the binomial approximation is sufficiently good for large pattern activity $k(n)$.

Here I have solved this problem. For this it was necessary to compute general expressions for the true potential distribution by defining different versions of the Willshaw–Palm probability including hetero-association, auto-association, and fixed and random pattern activities (see section 3). I then focused on the characterization of the probability distributions for random pattern activities. This involved computation of the raw and central moments of the Willshaw–Palm probability (section 4) from the corresponding moments of the binomial probability [25]. Finally, I have investigated the convergence of the two probabilities by determining conditions when the moments become identical. The analysis reveals that the moments become identical for almost any sublinear sparseness, for example, $k = O(n/(\log n)^2)$ (see section 5.4), and thus verifies the theory on associative networks for large pattern activities.

Appendix. Lemmas.

The following lemmas are required to prove the claims in this work. Proofs of the lemmas can be found in a technical report [23, 25] or in the standard literature of information theory, combinatorics, analysis, and probability theory (see, e.g., [9, 39]).

Let $X \in \{0, 1\}$ be a *binary* random variable with $p := \mathrm{pr}[X = 1]$ and information $I(p) := -p \,\mathrm{ld}\, p - (1 - p) \,\mathrm{ld}(1 - p)$. Further let Y be the result of transmitting X over a binary memoryless channel with transmission error probabilities $p_{01} := \mathrm{pr}[Y = 1|X = 0]$ and $p_{10} := \mathrm{pr}[Y = 0|X = 1]$. Then the *transinformation* between X and Y is

$$T(X;Y) = T(p, p_{01}, p_{10}) := I_Y(p, p_{01}, p_{10}) - I_{Y|X}(p, p_{01}, p_{10}), \tag{A.1}$$

where $I_Y(p, p_{01}, p_{10}) := I\left(p\,(1 - p_{10}) + (1 - p)\,p_{01}\right)$ is the information (or entropy) of Y and $I_{Y|X}(p, p_{01}, p_{10}) := p \cdot I(p_{10}) + (1 - p) \cdot I(p_{01})$ is the information of Y given X.

Now let $X \in \{0, 1, \ldots, N\}$ be a *binomially* distributed random variable with parameters N and P. Then X has expectation $E_{p_B} X = NP$, and the probability and moment generating functions are

$$\mathrm{pr}[X = x] = p_B(x; N, P) := \binom{N}{x} P^x (1-P)^{N-x}, \tag{A.2}$$

$$G_{p_B}(t; N, P) := E_{p_B} e^{tX} = (Pe^t + (1-P))^N. \tag{A.3}$$

Furthermore, substituting $Q := 1 - P$, it has been proven in [25] that the dth raw and central moments of X can be written as polynomials in Q,

$$\mathfrak{m}_{r,p_B}(d, N, P) := E_{p_B} X^d = \sum_{j=0}^{d} (-Q)^j \sum_{i=j}^{d} \binom{i}{j} S_{di} N^{\underline{i}} \quad \text{with} \tag{A.4}$$

$$E_{p_B}(X-\mu)^d = \sum_{j=0}^{d} (-Q)^j \binom{N}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (N-\mu-k)^d, \tag{A.5}$$

where $N^{\underline{i}} := N(N-1)\cdots(N-i+1)$ denotes a falling factorial, $S_{di} \geq 0$ are Stirling numbers of the second kind, and μ is an arbitrary offset. For $\mu = NP$, (A.5) yields the dth central moment $\mathfrak{m}_{c,p_B}(d, N, P)$ of the binomial probability. For $\mu = 0$, (A.5) becomes identical to the raw moment equation (A.4) [25]. The following lemma is the sieve formula of Sylvester and Poincaré:

$$\mathrm{pr}\left(\bigcup_{k=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathrm{pr}\left(\bigcap_{h=1}^{k} A_{i_h}\right). \tag{A.6}$$

The following combinatorial equations are true:

$$\binom{Y}{s-N} = \sum_{t=0}^{N} (-1)^{N+t} \binom{Y+t}{s} \binom{N}{t}, \tag{A.7}$$

$$\binom{n}{m}\binom{m}{p} = \binom{n}{p}\binom{n-p}{m-p} = \binom{n}{m-p}\binom{n-m+p}{p}, \tag{A.8}$$

$$\sum_{i=0}^{M} p_B(i; M, Q) \cdot (1-P)^{Ji} = (1 - Q(1-(1-P)^J))^M, \tag{A.9}$$

$$\binom{N}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (n-\mu-i)^d = N^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di} (N-\mu-j)^{\underline{i-j}}\ . \tag{A.10}$$

Equation (A.8) implies $B(a,b,c) = B(a,c,b)$ or $\binom{a}{b}\binom{a-b}{c} = \binom{a}{c}\binom{a-c}{b}$. Equation (A.9) is a variant of the binomial sum $(A+B)^M = \sum_{i=0}^{M} \binom{M}{i} A^i B^{M-i}$. For a proof of (A.10) see Lemma 3.1 in [25]. Here $N^{\underline{i}}$ denotes again a falling factorial, and $S_{di} \geq 0$ denotes Stirling numbers of the second kind.

Then we have used the bounds

$$(1-pq) \gtrless (1-p)^q \quad \text{for} \quad p \in (0;1) \quad \text{and} \quad q \substack{\in \\ \notin} (0;1), \tag{A.11}$$

$$1 + x \leq e^x \leq 1 + (e-1)x \quad \text{for} \quad 0 \leq x \leq 1, \tag{A.12}$$

where the first bound in (A.12) is true for any x. Finally, the following asymptotic equations are true for $n \to \infty$ with $|x(n)|, |y(n)| \to 0$:

$$e^x = 1 + x + \Theta(x^2), \qquad \ln(1+x) = x + \Theta(x^2), \tag{A.13}$$

$$e^{x+\Theta(y)} = 1 + x + \Theta(x^2) + \Theta(y), \qquad \frac{1}{1+x} = 1 - x + \Theta(x^2), \tag{A.14}$$

where for a function $f(n)$ we write $f(n) = \Theta(g(n))$ iff there are constants c_1, c_2, n_0 such that for any $n > n_0$ we have $c_1 g(n) < f(n) < c_2 g(n)$.

Acknowledgments. The author is grateful to Edgar Körner and Marc-Oliver Gewaltig for providing the opportunity to do this work at the Honda Research Institute, and also to Friedrich Sommer and Günther Palm for helpful discussions and comments.

REFERENCES

[1] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Information storage in neural networks with low levels of activity*, Phys. Rev. A, 35 (1987), pp. 2293–2303.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Statistical mechanics of neural networks near saturation*, Ann. Phys., 173 (1987), pp. 30–67.

[3] H. J. Bentz, M. Hagstroem, and G. Palm, *Information storage and effective data retrieval in sparse matrices*, Neural Networks, 2 (1989), pp. 289–293.

[4] H. Bosch and F. Kurfess, *Information storage capacity of incompletely connected associative memories*, Neural Networks, 11 (1998), pp. 869–876.

[5] V. Braitenberg, *Cell assemblies in the cerebral cortex*, in Theoretical Approaches to Complex Systems, R. Heim and G. Palm, eds., Lecture Notes in Biomath. 21, Springer-Verlag, Berlin, Heidelberg, New York, 1978, pp. 171–188.

[6] J. T. Buckingham, *Delicate Nets, Faint Recollections: A Study of Partially Connected Associative Network Memories*, Ph.D. thesis, University of Edinburgh, 1991.

[7] J. T. Buckingham and D. J. Willshaw, *Performance characteristics of associative nets*, Network: Computation in Neural Systems, 3 (1992), pp. 407–414.

[8] J. T. Buckingham and D. J. Willshaw, *On setting unit thresholds in an incompletely connected associative net*, Network: Computation in Neural Systems, 4 (1993), pp. 441–459.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[10] R. Fay, U. Kaufmann, A. Knoblauch, H. Markert, and G. Palm, *Integrating object recognition, visual attention, language and action processing on a robot using a neurobiologically motivated associative architecture*, in Proceedings of the NeuroRobotics Workshop at the 27th German GI Conference on Artificial Intelligence, University of Ulm, Germany, 2004.

[11] W. Gerstner and J. L. van Hemmen, *Associative memory in a network of "spiking" neurons*, Network, 3 (1992), pp. 139–164.

[12] B. Graham and D. Willshaw, *Improving recall from an associative memory*, Biol. Cybernet., 72 (1995), pp. 337–346.

[13] D. Greene, M. Parnas, and F. Yao, *Multi-index hashing for information retrieval*, in Proceedings of the 35th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, 1994, pp. 722–731.

[14] D. O. Hebb, *The Organization of Behavior. A Neuropsychological Theory*, Wiley, New York, 1949.

[15] R. Hecht-Nielsen, *Confabulation Theory*, Springer-Verlag, Heidelberg, 2007.

[16] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. USA, 79 (1982), pp. 2554–2558.

[17] P. Kanerva, *Sparse Distributed Memory*, MIT Press, Cambridge, MA, 1988.

[18] A. Knoblauch, *Optimal matrix compression yields storage capacity* 1 *for binary Willshaw associative memory*, in Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003, O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, eds., Lecture Notes in Comput. Sci. 2714, Springer-Verlag, Berlin, 2003, pp. 325–332.

[19] A. Knoblauch, *Synchronization and Pattern Separation in Spiking Associative Memory and Visual Cortical Areas*, Ph.D. thesis, Department of Neural Information Processing, University of Ulm, Germany, 2003.

[20] A. Knoblauch, *Neural associative memory for brain modeling and information retrieval*, Inform. Process. Lett., 95 (2005), pp. 537–544.

[21] A. Knoblauch, *Statistical implications of clipped Hebbian learning of cell assemblies*, Neurocomput., 65–66 (2005), pp. 647–652.

[22] A. Knoblauch, *On Compressing the Memory Structures of Binary Neural Associative Networks*, Internal Report HRI-EU 06-02, Honda Research Institute Europe, Offenbach/Main, Germany, 2006.

[23] A. Knoblauch, *Asymptotic Conditions for High-Capacity Neural Associative Networks*, Internal Report HRI-EU 07-02, Honda Research Institute Europe, Offenbach/Main, Germany, 2007.

[24] A. Knoblauch, *On the Computational Benefits of Inhibitory Neural Associative Networks*, Internal Report HRI-EU 07-05, Honda Research Institute Europe, Offenbach/Main, Germany, 2007.

[25] A. Knoblauch, *Closed-form expressions for the moments of the binomial probability distribution*, SIAM J. Appl. Math., 69 (2008), pp. 197–204.

[26] A. Knoblauch and G. Palm, *Pattern separation and synchronization in spiking associative memories and visual areas*, Neural Networks, 14 (2001), pp. 763–780.

[27] A. Knoblauch and G. Palm, *Scene segmentation by spike synchronization in reciprocally connected visual areas.* II. *Global assemblies and synchronization on larger space and time scales*, Biol. Cybernet., 87 (2002), pp. 168–184.

[28] T. Kohonen, *Associative Memory: A System Theoretic Approach*, Springer-Verlag, Berlin, 1977.

[29] P. E. Latham and S. Nirenberg, *Computing and stability in cortical networks*, Neural Comput., 16 (2004), pp. 1385–1412.

[30] D. Marr, *Simple memory: A theory for archicortex*, Philos. Trans. Roy. Soc. London Ser. B, 262 (1971), pp. 24–81.

[31] M. L. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge, MA, 1969.

[32] X. Mu, M. Artiklar, P. Watta, and M. H. Hassoun, *An RCE-based associative memory with application to human face recognition*, Neural Process. Lett., 23 (2006), pp. 257–271.

[33] J.-P. Nadal, *Associative memory: On the (puzzling) sparse coding limit*, J. Phys. A, 24 (1991), pp. 1093–1101.

[34] G. Palm, *On associative memories*, Biol. Cybernet., 36 (1980), pp. 19–31.

[35] G. Palm, *Neural Assemblies. An Alternative Approach to Artificial Intelligence*, Springer-Verlag, Berlin, 1982.

[36] G. Palm, *Computing with neural networks*, Science, 235 (1987), pp. 1227–1228.

[37] G. Palm, *Memory capacities of local rules for synaptic modification. A comparative review*, Concepts in Neurosci., 2 (1991), pp. 97–128.

[38] G. Palm and F. Sommer, *Associative data storage and retrieval in neural nets*, in Models of Neural Networks III, E. Domany, J. L. van Hemmen, and K. Schulten, eds., Springer-Verlag, New York, 1996, pp. 79–118.

[39] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1991.

[40] R. W. Prager and F. Fallside, *The modified Kanerva model for automatic speech recognition*, Computer Speech and Language, 3 (1989), pp. 61–81.

[41] D. A. Rachkovskij and E. M. Kussul, *Binding and normalization of binary sparse distributed representations by context-dependent thinning*, Neural Comput., 13 (2001), pp. 411–452.

[42] M. Rehn and F. T. Sommer, *Storing and restoring visual input with collaborative rank coding and associative memory*, Neurocomput., 69 (2006), pp. 1219–1223.

[43] F. T. Sommer and G. Palm, *Improved bidirectional retrieval of sparse patterns stored by Hebbian learning*, Neural Networks, 12 (1999), pp. 281–297.

[44] K. Steinbuch, *Die Lernmatrix*, Kybernetik, 1 (1961), pp. 36–45.

[45] H. Wersing and E. Körner, *Learning optimized features for hierarchical models of invariant object recognition*, Neural Comput., 15 (2003), pp. 1559–1588.

[46] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, *Non-holographic associative memory*, Nature, 222 (1969), pp. 960–962.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 197–204

CLOSED-FORM EXPRESSIONS FOR THE MOMENTS OF THE BINOMIAL PROBABILITY DISTRIBUTION[*]

ANDREAS KNOBLAUCH[†]

Abstract. This work develops closed-form expressions for the raw and central moments of the binomial probability distribution. For this I first derive a recursive formula for the raw moments from the moment generating function. Then it is shown that the recursion involved is essentially the same as for the Stirling numbers of the second kind. From this fact it is then possible to derive the closed formulae. Finally, I discuss an application of these formulae to the analysis of neural associative memory.

Key words. Stirling numbers, Pochhammer polynomial, neural associative memory, storage capacity, retrieval errors

AMS subject classifications. 05A15, 60C05, 92B20

DOI. 10.1137/070700024

1. The binomial probability and its moments. A random variable X is called *binomially distributed* with parameters n and p if the random variable takes value $x \in \{0, 1, 2, \ldots, n\}$ with probability

$$p_B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{1.1}$$

The *moment generating function* $G_B(s) := E_{p_B} e^{sX}$ of the binomial probability can then be computed using the binomial sum $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$,

$$G_B(s; n, p) = \sum_{x=0}^{n} \binom{n}{x} (pe^s)^x (1-p)^{n-x} = (pe^s + 1 - p)^n. \tag{1.2}$$

The dth *raw moment* $E_{p_B} X^d$ equals the dth derivative of the generating function $G_B(s)$ at $s = 0$ (e.g., [14]). For example, the mean value is $\mu := E_{p_B} X = n(pe^s + 1 - p)^{n-1} pe^s|_{s=0} = np$ and the second raw moment is $E_{p_B} X^2 = np((n-1)(pe^s + 1 - p)^{n-2} pe^s + (pe^s + 1 - p)^{n-1} e^s)|_{s=0} = np(np + 1 - p)$. Higher-order moments for larger d can be computed, in principle, by continuing this procedure, but computing higher-order derivatives of $G_B(s)$ becomes tedious with increasing d. In the following, we aim to find a recursive formula without referring to higher-order derivatives of $G_B(s)$ (see also [2] for a related approach).

2. Recursive formulae. To compute the higher-order derivatives of the moment generating function $G_B(s)$ for larger d, we can define auxiliary functions

$$H_d(s) := (pe^s)^d \text{ with derivative } H'_d(s) = dH_d(s), \tag{2.1}$$

$$F_d(s) := n^{\underline{d}} G_B(s; n-d, p) \text{ with derivative} \tag{2.2}$$

$$F'_d(s) = n^{\underline{d+1}} (pe^s + 1 - p)^{n-d-1} H_1(s) = F_{d+1}(s) H_1(s), \text{ and} \tag{2.3}$$

$$K_d(s) := F_d(s) H_d(s) \text{ with derivative} \tag{2.4}$$

$$K'_d(s) = F'_d(s) H_d(s) + dF_d(s) H_d(s) = K_{d+1}(s) + dK_d(s), \tag{2.5}$$

[*] Received by the editors August 13, 2007; accepted for publication (in revised form) May 28, 2008; published electronically September 8, 2008.
http://www.siam.org/journals/siap/69-1/70002.html

[†] Honda Research Institute Europe, Carl-Legien-Strasse 30, D-63073 Offenbach/Main, Germany (andreas.knoblauch@honda-ri.de).

where $n^{\underline{d}} = n(n-1)\cdots(n-d+1)$ denotes a falling factorial or Pochhammer symbol. Since $G_B(s) = K_0(s)$, we can obtain the higher-order derivatives of the moment generating function $G_B(s)$ recursively from (2.5), for example, $G_{p_B}^{(0)} = K_0$, $G_{p_B}^{(1)} = K_1$, $G_{p_B}^{(2)} = K_2 + K_1$, $G_{p_B}^{(3)} = K_3 + 2K_2 + K_2 + K_1 = K_3 + 3K_2 + K_1$. Thus, we can prove the following.

Lemma 2.1. *The dth derivative $G_B^{(d)}$ of the moment generating function $G_B(s)$ of the binomial probability $p_B(x; n, p)$ can be written as a weighted sum of functions $K_i(s)$:*

$$G_{p_B}^{(d)}(s) = \sum_{i=0}^{d} b_{di} K_i(s) \tag{2.6}$$

for appropriate coefficients b_{di}. The coefficients can be computed recursively from

$$b_{0i} = \delta_{i0}, \tag{2.7}$$

$$b_{di} = i b_{d-1,i} + b_{d-1,i-1}, \tag{2.8}$$

where δ_{ij} is the usual Kronecker symbol (1 for $i = j$, and 0 otherwise). For convenience, we further define $b_{di} = 0$ for $d < 0$, $i < 0$, or $i > d$.

Proof. Equation (2.7) follows from $G_B^{(0)} = K_0$ (see (2.4)). Equation (2.8) can then be shown inductively using (2.6) with (2.5):

$$\begin{aligned} G_{p_B}^{(d+1)}(s) &= \sum_{i=0}^{d} b_{di}(K_{i+1} + iK_i) = \sum_{i=1}^{d+1} b_{d,i-1} K_i + \sum_{i=0}^{d} b_{di} i K_i \\ &= \sum_{i=0}^{d+1} (i b_{di} + b_{d,i-1}) K_i. \qquad \square \end{aligned}$$

From this lemma and $K_i(0) = n^{\underline{i}} p^i$, we can give recursive formulae for the raw and central moments as summarized by the following theorem.

Theorem 2.2. *The dth raw and central moment of a binomially distributed random variable X with $\mathrm{pr}[X = x] = p_B(x; n, p)$, expectation $\mu := np$, and $q := 1 - p$ are*

$$E_{p_B} X^d = \sum_{i=0}^{d} b_{di} p^i n^{\underline{i}} \tag{2.9}$$

$$= \sum_{j=0}^{d} (-q)^j \sum_{i=j}^{d} \binom{i}{j} b_{di} n^{\underline{i}}, \tag{2.10}$$

$$E_{p_B}(X - \mu)^d = \sum_{i=0}^{d} \binom{d}{i} (-\mu)^{d-i} E_{p_B} X^i. \tag{2.11}$$

Equations (2.10), (2.11) can be obtained from the binomial sum (see section 1). Equation (2.10) is written as polynomial in q, which is useful for some applications (see section 5). The first few values of the coefficients b_{di} are shown in Table 2.1.

Table 2.1

Values of the binomial moment coefficients b_{di} for $0 \le d, i \le 10$. These coefficients can be used to compute the moments of the binomial probability (see (2.9)*) and are identical to the Stirling numbers of the second kind (see section* 3*).*

b_{di}	$i=0$	1	2	3	4	5	6	7	8	9	10
$d=0$	1	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0	0	0
3	0	1	3	1	0	0	0	0	0	0	0
4	0	1	7	6	1	0	0	0	0	0	0
5	0	1	15	25	10	1	0	0	0	0	0
6	0	1	31	90	65	15	1	0	0	0	0
7	0	1	63	301	350	140	21	1	0	0	0
8	0	1	127	966	1701	1050	266	28	1	0	0
9	0	1	255	3025	7770	6951	2646	462	36	1	0
10	0	1	511	9330	34105	42525	22827	5880	750	45	1

3. Relation to Stirling numbers of the second kind. The coefficients b_{di} for computing the binomial moments (2.9) are actually Stirling numbers of the second kind: The Stirling number of the second kind $S(d,i)$ is defined as the number of ways of partitioning a set of d elements into i nonempty sets, and one can show that $S(d,i)$ obeys the same recurrence relations (2.7), (2.8) as b_{di} (e.g., [1, 17, 9]). Closed formulae for the Stirling numbers of the second kind are well known, for example,

$$b_{di} = S(d,i) = \frac{(-1)^i}{i!} \sum_{k=0}^{i} (-1)^k \binom{i}{k} k^d. \tag{3.1}$$

Thus, inserting this into the formulae of Theorem 2.2 gives us already closed-form expressions for the moments of the binomial probability. However, these formulae can still be simplified using a generalization of the following generating function (e.g., see [17]):

$$n^d = \sum_{i=0}^{d} b_{di} n^{\underline{i}}. \tag{3.2}$$

The generalization is given by the following lemma.

Lemma 3.1.

$$\sum_{i=j}^{d} \binom{i}{j} b_{di} n^{\underline{i}} = \binom{n}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (n-k)^d. \tag{3.3}$$

Proof. Instead of (3.3), we prove the equivalent equation

$$C(d,j,n) := \sum_{i=j}^{d} i^{\underline{j}} b_{di} n^{\underline{i}} = n^{\underline{j}} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (n-k)^d \tag{3.4}$$

by induction over j. For $j=0$, the lemma is identical to (3.2). For larger j, we compute, using $ib_{di} = b_{d+1,i} - b_{d,i-1}$ (see (2.8)),

$$\begin{aligned} C(d,j+1,n) &= \sum_{i=j+1}^{d} i^{\underline{j+1}} b_{di} n^{\underline{i}} = \sum_{i=j+1}^{d} i^{\underline{j}} i b_{di} n^{\underline{i}} - j \sum_{i=j+1}^{d} i^{\underline{j}} b_{di} n^{\underline{i}} \\ &= \sum_{i=j+1}^{d} i^{\underline{j}} b_{d+1,i} n^{\underline{i}} - \sum_{i=j+1}^{d} i^{\underline{j}} b_{d,i-1} n^{\underline{i}} - j \sum_{i=j+1}^{d} i^{\underline{j}} b_{di} n^{\underline{i}}. \end{aligned}$$

The three sums can be written individually:

$$\begin{aligned}
S_1 &:= \sum_{i=j+1}^{d} i^{\underline{j}} b_{d+1,i} n^{\underline{i}} = C(d+1,j,n) - j! b_{d+1,j} n^{\underline{j}} - (d+1)^{\underline{j}} n^{\underline{d+1}},\\
S_2 &:= \sum_{i=j+1}^{d} i^{\underline{j}} b_{d,i-1} n^{\underline{i}}\\
&= n \sum_{i=j+1}^{d} (i-1)^{\underline{j}} b_{d,i-1} (n-1)^{\underline{i-1}} + jn \sum_{i=j+1}^{d} (i-1)^{\underline{j-1}} b_{d,i-1} (n-1)^{\underline{i-1}}\\
&= n \sum_{i=j}^{d-1} i^{\underline{j}} b_{d,i} (n-1)^{\underline{i}} + jn \sum_{i=j}^{d-1} i^{\underline{j-1}} b_{d,i} (n-1)^{\underline{i}}\\
&= nC(d,j,n-1) - nd^{\underline{j}}(n-1)^{\underline{d}}\\
&\quad + jnC(d,j-1,n-1) - jn(j-1)^{\underline{j-1}} b_{d,j-1}(n-1)^{\underline{j-1}} - jnd^{\underline{j-1}}(n-1)^{\underline{d}},\\
S_3 &:= j \sum_{i=j+1}^{d} i^{\underline{j}} b_{di} n^{\underline{i}} = jC(d,j,n) - jj^{\underline{j}} b_{dj} n^{\underline{j}},
\end{aligned}$$

where we used $b_{dd} = 1$ for $d \geq 0$. For the second sum, S_2, we used $i^{\underline{j}} = ((i-j)+j)(i-1)^{\underline{j-1}}$. Fortunately, in $S_1 - S_2 - S_3$ all the non-C terms cancel out: The b terms cancel out because with $b_{d+1,j} = jb_{d,j} + b_{d,j-1}$ from (2.8), we have

$$-j!n^{\underline{j}} b_{d+1,j} + j!n^{\underline{j}} b_{d,j-1} + jj!n^{\underline{j}} b_{dj} = j!n^{\underline{j}}(-b_{d+1,j} + b_{d,j-1} + jb_{dj}) = 0.$$

The remaining non-C and non-b terms cancel out because

$$-(d+1)^{\underline{j}} n^{\underline{d+1}} + d^{\underline{j}} n^{\underline{d+1}} + jd^{\underline{j-1}} n^{\underline{d+1}} = n^{\underline{d+1}} d^{\underline{j-1}}(-(d+1) + (d-j+1) + j) = 0.$$

Thus, using the induction hypothesis, we have simply

$$\begin{aligned}
C(d,j+1,n) &= C(d+1,j,n) - nC(d,j,n-1) - jnC(d,j-1,n-1) - jC(d,j,n)\\
&= n^{\underline{j}} \sum_{k=0}^{j} \binom{j}{k} (-1)^k (n-k)^{d+1}\\
&\quad - n(n-1)^{\underline{j}} \sum_{k=0}^{j} \binom{j}{k} (-1)^k (n-(k+1))^d\\
&\quad - jn(n-1)^{\underline{j-1}} \sum_{k=0}^{j-1} \binom{j-1}{k} (-1)^k (n-(k+1))^d\\
&\quad - jn^{\underline{j}} \sum_{k=0}^{j} \binom{j}{k} (-1)^k (n-k)^d\\
&= \sum_{k=0}^{j+1} a_k (n-k)^d.
\end{aligned}$$

In the last line we have simply summed over the $(n-k)^d$ terms. Thus, our proof is finished if we can show that $a_k = n^{\underline{j+1}}\binom{j+1}{k}(-1)^k$ for $k = 0, 1, \ldots, j+1$. The highest coefficient a_{j+1} gets contributions only from the second sum:

$$a_{j+1} = -n^{\underline{j+1}}(-1)^j(n-(j+1))^d = n^{\underline{j+1}}\binom{j+1}{j+1}(-1)^k.$$

The lowest coefficient a_0 gets contributions only from the first and fourth sums:

$$a_0 = n^{\underline{j}}n - jn^{\underline{j}} = n^{\underline{j+1}} = n^{\underline{j+1}}\binom{j+1}{0}(-1)^0.$$

The remaining intermediary coefficients a_k for $k = 1, 2, \dots, j$ get contributions from all four sums:

$$\begin{aligned} a_k &= n^{\underline{j}}\binom{j}{k}(-1)^k(n-k) - n^{\underline{j+1}}\binom{j}{k-1}(-1)^{k-1} \\ &\quad - jn^{\underline{j}}\binom{j-1}{k-1}(-1)^{k-1} - jn^{\underline{j}}\binom{j}{k}(-1)^k \\ &= (-1)^k n^{\underline{j}}\binom{j}{k}\left((n-k)+(n-j)\frac{k}{j-k+1}+j\frac{k}{j}-j\right) \\ &= (-1)^k n^{\underline{j}}\binom{j}{k}\frac{(n-j)(j+1)}{j-k+1} = n^{\underline{j+1}}\binom{j+1}{k}(-1)^k. \end{aligned}$$

Thus, we have proven (3.4). □

A useful variant of Lemma 3.1, including an offset μ, is

$$\binom{n}{j}\sum_{k=0}^{j}(-1)^k\binom{j}{k}(n-\mu-k)^d = n^{\underline{j}}\sum_{i=j}^{d}\binom{i}{j}b_{di}(n-\mu-j)^{\underline{i-j}}. \tag{3.5}$$

4. Closed formulae. The following theorem summarizes the main results of this work.

THEOREM 4.1. *Let X be a binomially distributed random variable with probability function $p_B(x; n, p)$ (see* (1.1)*). Further, let $q := 1 - p$ and let b_{di} be Stirling numbers of the second kind (see Table* 2.1 *and* (2.7), (2.8), *and* (3.1)*). Then the dth raw moment of the binomial probability p_B can be written as*

$$E_{p_B}X^d = \sum_{i=0}^{d} b_{di}p^i n^{\underline{i}} \tag{4.1}$$

$$= \sum_{i=0}^{d}(-p)^i\binom{n}{i}\sum_{k=0}^{i}(-1)^k\binom{i}{k}k^d \tag{4.2}$$

$$= \sum_{j=0}^{d}(-q)^j\sum_{i=j}^{d}\binom{i}{j}b_{di}n^{\underline{i}} \tag{4.3}$$

$$= \sum_{j=0}^{d}(-q)^j\binom{n}{j}\sum_{k=0}^{j}(-1)^k\binom{j}{k}(n-k)^d. \tag{4.4}$$

For the dth raw moment, the following bounds are true:

$$(np)^d \le E_{p_B}X^d \le n^d. \tag{4.5}$$

For an arbitrary offset μ, we have

$$E_{p_B}(X-\mu)^d = \sum_{i=0}^{d} \binom{d}{i} (-\mu)^{d-i} E_{p_B} X^i \tag{4.6}$$

$$= \sum_{j=0}^{d} (-p)^j \binom{n}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (k-\mu)^d \tag{4.7}$$

$$= \sum_{j=0}^{d} (-q)^j n^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} b_{di} (n-\mu-j)^{\underline{i-j}} \tag{4.8}$$

$$= \sum_{j=0}^{d} (-q)^j \binom{n}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (n-\mu-k)^d. \tag{4.9}$$

In particular, for $\mu = E_{p_B}X = np$, we obtain the dth central moment of the binomial probability p_B. For $z - d \geq \mu \geq E_{p_B}X$, the following bound is true:

$$|E_{p_B}(X-\mu)^d| \leq \sum_{j=0}^{d} \sum_{i=j}^{d} \binom{i}{j} b_{di} (nq)^i \quad (\sim (nq)^d \; for \; fixed \; d \; and \; nq \to \infty). \tag{4.10}$$

Proof. Equations (4.1), (4.2) follow from (2.9), (3.1). Equations (4.3), (4.4) follow from (2.10), (3.3). The bounds of (4.5) follow simply from $p^d \leq p^i \leq 1$ and (3.2) because (4.1) is obviously a sum of nonnegative numbers. Equation (4.6) is (2.11). Equation (4.7) follows by inserting (4.2) into (4.6) and applying the binomial sum (see section 1). Similarly, (4.9) follows from inserting (4.4) into (4.6). Equation (4.8) follows from (4.9) with (3.5). Equation (4.10) follows from (4.8) because $n^{\underline{j}} \leq n^j$ and $(n-\mu-j)^{\underline{i-j}} \leq (nq)^{i-j}$ for $n \geq \mu + d$ and $\mu \geq E_{p_B}X = np$. □

5. Related work and application to the analysis of neural associative networks. Computing the higher-order moments of a binomially distributed random variable is rarely emphasized. Standard textbooks give expressions for the moment generating function (1.2) and some lower-order moments such as mean, variance, and, perhaps, skewness and kurtosis, but higher-order moments are usually neglected (e.g., see [14, 16]). For some applications it may be sufficient to approximate a binomial random variable by either a Gaussian or a Poissonian where closed-form expressions for higher-order moments are known. For example, for large variance $np(1-p) \to \infty$, according to the DeMoivre–Laplace theorem, the binomial probability becomes similar to a Gaussian with the same mean and variance. Likewise, for $n \to \infty$ and finite $np \to \lambda < \infty$, the binomial becomes Poissonian. However, for applications as described below, these approximations are not appropriate, and it is necessary to find an exact formula.

A previous attempt [2] to compute the higher-order moments of the binomial distribution revealed recursive expressions similar to those developed in section 2, but was restricted to the special case $p = 0.5$. Moreover, the recursive form was not appropriate for efficient computation or application in further analyses. In contrast to [2], this work provides general recursive and nonrecursive (or closed-form) expressions for the higher-order moments of the binomial distribution.

My main motivation to obtain a closed formula for the binomial moments comes from analyzing storage capacity and retrieval error probabilities in neural associative memory networks [21, 12, 4, 18, 6]. *Associative memories* are systems that contain information about a finite set of associations between pattern vector pairs $\{(\mathbf{u}^{\mu} \mapsto \mathbf{v}^{\mu}) : \mu = 1, \ldots, M\}$ [10]. Given a possibly noisy address pattern $\tilde{\mathbf{u}}$, the problem is to find a target pattern $\mathbf{v}^{\mu}$ for which the corresponding address pattern $\mathbf{u}^{\mu}$ is most similar to $\tilde{\mathbf{u}}$.

Neural associative networks have wide applications for both artificial intelligence (e.g., visual object recognition [10, 15]) and modeling the brain (e.g., [13, 6, 22, 5, 19, 20]). In neural implementations the associations are stored in a matrix $\mathbf{A}$ describing the synaptic connections between two cell populations u and v. Here the retrieval result $\hat{\mathbf{v}}^{\mu}$ may differ from the original pattern $\mathbf{v}^{\mu}$. This is due to retrieval noise being an increasing function of the memory load or the number of stored associations. In general, the probability of a retrieval error can be computed from the neuron potential distribution as obtained by propagating the address pattern $\tilde{\mathbf{u}}$ through the synaptic matrix $\mathbf{A}$.

One of the most efficient models is the so-called *Willshaw network* with binary neurons and synapses [21]. Here the synaptic matrix is simply $\mathbf{A} = \vee_{\mu=1}^{M} \mathbf{u}^{\mu,\mathbf{T}} \mathbf{v}^{\mu}$, and the retrieval error probabilities can be computed from the so-called Willshaw–Palm distribution of neuron potentials $\mathbf{x} = \tilde{\mathbf{u}}^{\mathbf{T}} \mathbf{A}$. Since the Willshaw–Palm distribution is more difficult to formulate, many analyses of neural associative memory actually rely on a binomial approximation (e.g., [21, 12, 11, 3, 18]). However, it is unclear for which network parameters this approximation is sufficiently accurate. In a further paper [8] (see also [7]), I will compute the moments of the Willshaw–Palm distribution from the binomial moments. For this it is sufficient to replace q^j in (4.4), (4.9) by some more complex term $q^{(j)}$. With this it will be possible to compare the exact potential distribution to the binomial approximation and determine asymptotic conditions when they become identical.

Acknowledgments. I am grateful to Edgar Körner and Marc-Oliver Gewaltig for providing the opportunity to do this work at the Honda Research Institute. I am also grateful to Günther Palm and Friedrich Sommer for helpful discussions and comments.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed., Dover, New York, 1972.

[2] A. Benyi and S. M. Manago, *A recursive formula for moments of a binomial distribution*, College Math. J., 36 (2005), pp. 68–72.

[3] H. Bosch and F. Kurfess, *Information storage capacity of incompletely connected associative memories*, Neural Networks, 11 (1998), pp. 869–876.

[4] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. USA, 79 (1982), pp. 2554–2558.

[5] R. A. Jortner, S. S. Farivar, and G. Laurent, *A simple connectivity scheme for sparse coding in an olfactory system*, J. Neurosci., 27 (2007), pp. 1659–1669.

[6] A. Knoblauch, *Neural associative memory for brain modeling and information retrieval*, Inform. Process. Lett., 95 (2005), pp. 537–544.

[7] A. Knoblauch, *Asymptotic Conditions for High-Capacity Neural Associative Networks*, Internal Report HRI-EU 07-02, Honda Research Institute Europe GmbH, Offenbach/Main, Germany, 2007.

[8] A. Knoblauch, *Neural associative memory and the Willshaw–Palm probability distribution*, SIAM J. Appl. Math., 69 (2008), pp. 169–196.

[9] D. E. Knuth, *The Art of Computer Programming, Volume* 1: *Fundamental Algorithms*, 3rd ed., Addison-Wesley, Reading, MA, 1997.

[10] T. Kohonen, *Associative Memory: A System-Theoretical Approach*, Springer, Berlin, 1977.

[11] J.-P. Nadal, *Associative memory: On the (puzzling) sparse coding limit*, J. Phys. A, 24 (1991), pp. 1093–1101.

[12] G. Palm, *On associative memories*, Biol. Cybern., 36 (1980), pp. 19–31.

[13] G. Palm, *Neural Assemblies. An Alternative Approach to Artificial Intelligence*, Springer, Berlin, 1982.

[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw–Hill, New York, 1991.

[15] M. Rehn and F. T. Sommer, *Storing and restoring visual input with collaborative rank coding and associative memory*, Neurocomputing, 69 (2006), pp. 1219–1223.

[16] J. Riordan, *An Introduction to Combinatorial Analysis*, Wiley, New York, 1958.

[17] S. Roman, *The Umbral Calculus*, Pure Appl. Math. 111, Academic Press, New York, 1984.

[18] F. T. Sommer and G. Palm, *Improved bidirectional retrieval of sparse patterns stored by Hebbian learning*, Neural Networks, 12 (1999), pp. 281–297.

[19] L. G. Valiant, *Memorization and association on a realistic neural model*, Neural Comput., 17 (2005), pp. 527–555.

[20] L. G. Valiant, *A quantitative theory of neural computation*, Biol. Cybern., 95 (2006), pp. 205–211.

[21] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, *Nonholographic associative memory*, Nature, 222 (1969), pp. 960–962.

[22] X. Xie, R. H. R. Hahnloser, and H. S. Seung, *Selectively grouping neurons in recurrent networks of lateral inhibition*, Neural Comput., 14 (2002), pp. 2627–2646.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 205–250

ABSOLUTE STABILITY AND COMPLETE SYNCHRONIZATION IN A CLASS OF NEURAL FIELDS MODELS*

OLIVIER FAUGERAS†, FRANÇOIS GRIMBERT†, AND JEAN-JACQUES SLOTINE‡

Abstract. Neural fields are an interesting option for modeling macroscopic parts of the cortex involving several populations of neurons, like cortical areas. Two classes of neural field equations are considered: voltage- and activity-based. The spatio-temporal behavior of these fields is described by nonlinear integro-differential equations. The integral term, computed over a compact subset of $\mathbb{R}^q$, $q = 1, 2, 3$, involves space and time varying, possibly nonsymmetric, intracortical connectivity kernels. Contributions from white matter afferents are represented as external input. Sigmoidal nonlinearities arise from the relation between average membrane potentials and instantaneous firing rates. Using methods of functional analysis, we characterize the existence and uniqueness of a solution of these equations for general, homogeneous (i.e., independent of the spatial variable), and spatially locally homogeneous inputs. In all cases we give sufficient conditions on the connectivity functions for the solutions to be absolutely stable, that is to say, asymptotically independent of the initial state of the field. These conditions bear on some compact operators defined from the connectivity kernels, the maximal slope of the sigmoids, and the time constants used in describing the temporal shape of the postsynaptic potentials. Numerical experiments are presented to illustrate the theory. To our knowledge this is the first time that such a complete analysis of the problem of the existence and uniqueness of a solution of these equations has been obtained. Another important contribution is the analysis of the absolute stability of these solutions—more difficult but more general than the linear stability analysis which it implies. The reason we have been able to complete this work is our use of the functional analysis framework and the theory of compact operators in a Hilbert space which has allowed us to provide simple mathematical answers to some of the questions raised by modelers in neuroscience.

Key words. neural fields, integro-differential equations, compact operators, Hilbert space, absolute stability, complete synchronization, Lyapunov function, neural masses, cortical columns

AMS subject classifications. 34G20, 34L30, 47B15, 47G10, 47G20, 47J05, 82C32, 92B20, 92C20

DOI. 10.1137/070694077

1. Introduction.

We model neural fields as continuous networks of cortical units and investigate the ability of these units to completely synchronize, i.e., to produce the same output when receiving the same input independently of their initial state. We therefore emphasize the dynamics and the spatio-temporal behavior of these networks.

Cortical units are built from a local description of the dynamics of a number of interacting neuron populations, called *neural masses* [15], where the spatial structure of the connections is neglected. These “vertically” built units can be thought of as *cortical columns* [29, 30, 3]. Probably the most well-known neural mass–based column model is that of Jansen and Rit [21] based on the original work of da Silva, Van Rotterdam, and colleagues [25, 26, 38]. A complete analysis of the bifurcations of this model can be found in [18]. More realistic models can be derived from experimental connectivity studies, such as the one shown in Figure 1. This figure, adapted

*Received by the editors June 8, 2007; accepted for publication (in revised form) May 22, 2008; published electronically September 19, 2008. This work was partially supported by Elekta Instrument AB, the “Fondation d’Entreprise EADS,” and the EC IP project #015879 FACETS.

http://www.siam.org/journals/siap/69-1/69407.html

†INRIA/ENPC/ENS, Odyssée Team, 2004 route des Lucioles, Sophia-Antipolis, France (olivier.faugeras@sophia.inria.fr, francois.grimbert@sophia.inria.fr).

‡Nonlinear Systems Laboratory, MIT, 77 Massachusetts Avenue, Cambridge, MA (jjs@mit.edu).

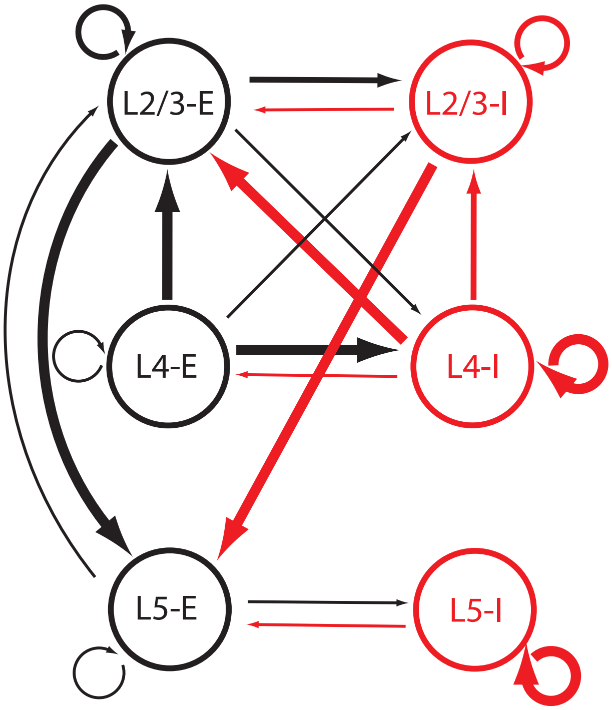


Fig. 1. *A simplified model of local cortical interactions based on six neuron populations. This local connectivity graph can be seen as a model of a cortical column composed of six interacting neural masses. There are three layers corresponding to cortical layers* II/III, IV, *and* V *and two types of neurons (excitatory ones in black and inhibitory ones in red) in each of these layers. The size of the arrows gives an idea of the average strength of the postsynaptic potentials elicited by the presynaptic neurons; see section* 2.1.1. *This figure is adapted from* [19].

from [19], is based on the work of Thomson and Bannister [37]. It shows the local connectivity graph of six populations of neurons and can be thought of as a model of a column comprising six interacting neural masses.

Such columns are then assembled spatially to form the neural field, which is meant to represent a macroscopic part of the neocortex, e.g., a visual area such as V1. Connections between columns are intracortical (gray matter) connections. Connections made via white matter with, e.g., such visual areas as the LGN or V2 are also considered in our models but are treated as input/output quantities.

There are at least three reasons why we think this is the relevant granularity to do modeling:

- Realistic modeling of a macroscopic part of the brain at the scale of the neuron is still difficult for obvious complexity reasons. Starting from mesoscopic building blocks like neural masses, described by the average activity of their neurons, is therefore a reasonable choice.
- While MEG and scalp EEG recordings mostly give a bulk signal of a cortical area, multielectrode recordings, in vitro experiments on pharmacologically treated brain slices, and new imaging techniques like extrinsic optical imaging can provide a spatially detailed description of neural masses dynamics in a macroscopic part of the brain like an area.
- The column/area scales correspond to available local connectivity data. In-

deed, these are obtained by averaging over local populations of neurons we can think of as neural masses. Besides, local (intracolumnar) connectivity is supposed to be spatially invariant within an area.

We now present a general mathematical framework for neural field modeling that agrees with the ideas of using average descriptions of neuronal activity and spatial invariance of the local connectivity across the field. This framework uses the elegant tools of functional analysis with the advantage of providing simple characterizations of some important properties of neural field equations.

In section 2 we describe the local and spatial models of neural masses and derive the equations that govern their spatio-temporal variations. In section 3 we analyze the problem of the existence and uniqueness of the smooth general and homogeneous solutions of these equations. In section 4 we study the absolute stability of these solutions, i.e., their robustness to arbitrary perturbations caused by changes of the initial conditions. In section 5 we extend this analysis to the absolute stability of the homogeneous, i.e., independent of space, solutions when they exist. A consequence of the absolute stability is the ability of the network to completely synchronize. In section 6 we revisit the functional framework of our analysis and extend our results to nonsmooth functions with the effect that we can discuss the existence and absolute stability of locally homogeneous solutions. We also propose another extension of the model by generalizing the previous results to higher order synaptic responses. In section 7 we present a number of numerical experiments to illustrate the theory and conclude in section 8.

2. The models.

We discuss local and spatial models.

2.1. The local models.

We consider n interacting populations of neurons such as those shown in Figure 1. The following derivation is built after Ermentrout's review [10]. We consider that each neural population i is described by its average membrane potential $V_i(t)$ or by its average instantaneous firing rate $\nu_i(t)$, the relation between the two quantities being of the form $\nu_i(t) = S_i(V_i(t))$ [16, 8], where S_i is sigmoidal. The functions S_i, $i = 1, \dots, n$, satisfy the properties introduced in the following definition.

DEFINITION 2.1. *For all $i = 1, \dots, n$, S_i and S_i' are positive and bounded (S_i' is the derivative of the function S_i). We note $S_{im} = \sup_x S_i(x)$, $S_m = \max_i S_{im}$, $S_{im}' = \sup_x S_i'(x)$, and $DS_m = \max_i S_{im}'$. Finally, we define $D\mathbf{S}_m$ as the diagonal matrix* $\mathrm{diag}(S_{im}')$.

Neurons in population j are connected to neurons in population i. A single action potential from neurons in population j is seen as a postsynaptic potential $PSP_{ij}(t-s)$ by neurons in population i, where s is the time of the spike hitting the synapse and t the time after the spike. We neglect the delays due to the distance traveled down the axon by the spikes.

Assuming that the postsynaptic potentials sum linearly, the average membrane potential of population i is

$$V_i(t) = \sum_{j,k} PSP_{ij}(t - t_k),$$

where the sum is taken over the arrival times of the spikes produced by the neurons in population j. The number of spikes arriving between t and $t+dt$ is $\nu_j(t)dt$. Therefore, we have

$$V_i(t) = \sum_j \int_{t_0}^{t} PSP_{ij}(t-s)\nu_j(s)\,ds = \sum_j \int_{t_0}^{t} PSP_{ij}(t-s)S_j(V_j(s))\,ds$$

or, equivalently,

$$\nu_i(t) = S_i\left(\sum_j \int_{t_0}^{t} PSP_{ij}(t-s)\nu_j(s)\,ds\right). \tag{2.1}$$

The PSP_{ij}s can depend on several variables in order to account for adaptation, learning, etc.

There are two main simplifying assumptions that appear in the literature [10], and they yield two different models.

2.1.1. The voltage-based model. The assumption, made in [20], is that the postsynaptic potential has the same shape no matter which presynaptic population caused it; the sign and amplitude may vary, though. This leads to the relation

$$PSP_{ij}(t) = W_{ij}PSP_i(t).$$

PSP_i represents the unweighted shape of the postsynaptic potentials and W_{ij} is the average strength of the postsynaptic potentials elicited by neurons of type j on neurons of type i. In biophysical connectivity models, like the one presented in Figure 1, the W_{ij}s should be chosen proportionally to the number of presynaptic cells, the average amplitude of postsynaptic potentials, and the probability of connection between the considered neuron species [17, 13]. In particular, if $W_{ij} > 0$, the population j excites population i, whereas it inhibits it when $W_{ij} < 0$.

Finally, if we assume that $PSP_i(t) = e^{-t/\tau_i}Y(t)$ (where Y is the Heaviside distribution) or, equivalently, that

$$\tau_i \frac{dPSP_i(t)}{dt} + PSP_i(t) = \tau_i\delta(t), \tag{2.2}$$

we end up with the system of ordinary first order differential equations

$$\frac{dV_i(t)}{dt} + \frac{V_i(t)}{\tau_i} = \sum_j W_{ij}S_j(V_j(t)) + I^i_{\mathrm{ext}}(t), \tag{2.3}$$

which describes the dynamic behavior of a cortical column. We have added an external current $I_{\mathrm{ext}}(t)$ to model the nonlocal connections of population i.

The approach developed in this article generalizes easily to the case of more sophisticated postsynaptic potentials models resulting in higher order differential equations, as shown in section 6.3.

We introduce the $n \times n$ matrix $\mathbf{W} = (W_{ij})_{i,j}$ and the function $\mathbf{S}$: $\mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{S}(\mathbf{x})$ is the vector of coordinates $S_i(x_i)$. We rewrite (2.3) in vector form and obtain the system of n ordinary differential equations

$$\mathbf{V}' = -\mathbf{L}\mathbf{V} + \mathbf{W}\mathbf{S}(\mathbf{V}) + \mathbf{I}_{\mathrm{ext}}, \tag{2.4}$$

where $\mathbf{L}$ is the diagonal matrix $\mathbf{L} = \mathrm{diag}(1/\tau_{\mathrm{i}})$.

2.1.2. The activity-based model.

The assumption is that the shape of a postsynaptic potential depends only on the nature of the presynaptic cell, that is,

$$PSP_{ij}(t) = W_{ij} PSP_j(t).$$

As above, we suppose that $PSP_i(t)$ satisfies the differential equation (2.2) and define the activity to be

$$A_j(t) = \int_{t_0}^{t} PSP_j(t-s)\nu_j(s)\, ds.$$

A similar derivation yields the following set of n ordinary differential equations:

$$\frac{dA_i(t)}{dt} + \frac{A_i(t)}{\tau_i} = S_i\left(\sum_j W_{ij} A_j(t) + I^i_{\text{ext}}(t)\right), \quad i = 1, \dots, n.$$

We rewrite this in vector form as

$$\mathbf{A}' = -\mathbf{L}\mathbf{A} + \mathbf{S}(\mathbf{W}\mathbf{A} + \mathbf{I}_{\text{ext}}). \tag{2.5}$$

We introduce the following definition.

Definition 2.2. *We denote $\tau_{\max}$ the maximum of the decay time constants τ_i, $i = 1, \dots, n$:*

$$\tau_{\max} = \max_i \tau_i.$$

2.2. Neural fields models.

We now combine these local models to form a continuum of columns, e.g., in the case of a model of a significant part Ω of the cortex. From now on we consider a compact subset Ω of $\mathbb{R}^q$, $q = 1, 2, 3$. This encompasses several cases of interest.

When $q = 1$ we deal with one-dimensional neural fields. Even though this appears to be of limited biological interest, it is one of the most widely studied cases because of its relative mathematical simplicity and because of the insights one can gain on the more realistic situations.

When $q = 2$ we discuss properties of two-dimensional neural fields. This is perhaps more interesting from a biological point of view since Ω can be viewed as a piece of cortex where the third dimension, its thickness, is neglected. This case has received far less attention than the previous one, probably because of the increased mathematical difficulty.

Finally, $q = 3$ allows us to discuss properties of volumes of neural masses, e.g., cortical sheets where their thickness is taken into account [22, 4].

The results that are presented in this paper are independent of q. Nevertheless, we have a good first approximation of a real cortical area with $q = 2$, and cortical depth given by the index $i = 1, \dots, n$ of the considered cortical population, following the idea of a field composed of columns or, equivalently, of interconnected cortical layers.

We denote $\mathbf{V}(\mathbf{r}, t)$ (resp., $\mathbf{A}(\mathbf{r}, t)$) the n-dimensional state vector at the point $\mathbf{r}$ of the continuum and at time t. We introduce the $n \times n$ matrix function $\mathbf{W}(\mathbf{r}, \mathbf{r}', t)$, which describes how the neural mass at point $\mathbf{r}'$ influences that at point $\mathbf{r}$ at time t. More precisely, $W_{ij}(\mathbf{r}, \mathbf{r}', t)$ describes how population j at point $\mathbf{r}'$ influences population i at point $\mathbf{r}$ at time t. We call $\mathbf{W}$ the connectivity matrix function. Neglecting, as in the local case above, the delays due to the distance between the neural masses, we extend (2.4) to

$$\mathbf{V}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{V}(\mathbf{r},t) + \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{S}(\mathbf{V}(\mathbf{r}',t))\,d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t) \tag{2.6}$$

and (2.5) to

$$\mathbf{A}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{A}(\mathbf{r},t) + \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{A}(\mathbf{r}',t)\,d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t)\right). \tag{2.7}$$

$\mathbf{V}_t$ (resp., $\mathbf{A}_t$) stands for the partial derivative of the vector $\mathbf{V}$ (resp., $\mathbf{A}$) with respect to the time variable t. A special case which will be considered later is when $\mathbf{W}$ is translation invariant: $\mathbf{W}(\mathbf{r},\mathbf{r}',t) = \mathbf{W}(\mathbf{r}-\mathbf{r}',t)$. We give below sufficient conditions on $\mathbf{W}$ and $\mathbf{I}_{\text{ext}}$ for (2.6) and (2.7) to be well defined and study their solutions.

3. Existence and uniqueness of a solution. In this section we deal with the problem of the existence and uniqueness of a solution to (2.6) and (2.7) for a given set of initial conditions. Unlike previous authors [12, 5, 28], we consider the case of a neural field with the effect that we have to use the tools of functional analysis to characterize their properties.

We start with the assumption that the state vectors $\mathbf{V}$ and $\mathbf{A}$ are differentiable (resp., continuous) functions of the time (resp., the space) variable. This is certainly reasonable in terms of the temporal variations because we are essentially modeling large populations of neurons and do not expect to be able to represent time transients. It is far less reasonable in terms of the spatial dependency since one should allow neural mass activity to be spatially distributed in a locally nonsmooth fashion with areas of homogeneous cortical activity separated by smooth boundaries. A more general assumption is proposed in section 6. But it turns out that most of the groundwork can be done in the setting of continuous functions.

Let $\mathcal{F}$ be the set $\mathbf{C}_n(\Omega)$ of the continuous functions from Ω to $\mathbb{R}^n$. This is a Banach space for the norm $\|\mathbf{V}\|_{n,\infty} = \max_{1\le i\le n}\sup_{\mathbf{r}\in\Omega}|\mathbf{V}_i(\mathbf{r})|$; see section A.1. We denote by J a closed interval of the real line containing 0.

We will need the following lemma several times.

Lemma 3.1. *We have the following inequalities for all* $\mathbf{x},\mathbf{y}\in\mathcal{F}$ *and* $\mathbf{r}'\in\Omega$*:*

$$\|\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{S}(\mathbf{y}(\mathbf{r}'))\|_\infty \le DS_m\|\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')\|_\infty \quad \text{and}$$
$$\|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{y})\|_{n,\infty} \le DS_m\|\mathbf{x}-\mathbf{y}\|_{n,\infty}.$$

Proof. $\mathbf{S}$ is smooth so we can perform a zeroth-order Taylor expansion with integral remainder [9] and write

$$\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{S}(\mathbf{y}(\mathbf{r}')) = \left(\int_0^1 D\mathbf{S}(\mathbf{y}(\mathbf{r}') + \zeta(\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')))\,d\zeta\right)(\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')),$$

and, because of Lemma A.1 and Definition 2.1

$$\begin{aligned}\|\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{S}(\mathbf{y}(\mathbf{r}'))\|_\infty &\le \int_0^1 \|D\mathbf{S}(\mathbf{y}(\mathbf{r}') + \zeta(\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')))\|_\infty\,d\zeta\,\|\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')\|_\infty \\ &\le DS_m\|\mathbf{x}(\mathbf{r}')-\mathbf{y}(\mathbf{r}')\|_\infty.\end{aligned}$$

This proves the first inequality. The second follows immediately. □

3.1. General solution. A function $\mathbf{V}(t)$ is thought of as a mapping $\mathbf{V} : \mathrm{J} \rightarrow \mathcal{F}$. This means that $\mathbf{V}(t)$ is now a function defined in Ω. Equations (2.6) and (2.7) are formally recast as an initial value problem (see, e.g., [11]):

$$\left\{ \begin{array}{rcl} \mathbf{V}'(t) &=& f(t, \mathbf{V}(t)), \\ \mathbf{V}(0) &=& \mathbf{V}_0, \end{array} \right. \tag{3.1}$$

where $\mathbf{V}_0$ is an element of $\mathcal{F}$ and the function f from $\mathrm{J} \times \mathcal{F}$ is equal to f_v defined by the right-hand side of (2.6):

$$f_v(t, \mathbf{x})(\mathbf{r}) = -\mathbf{L}\mathbf{x}(\mathbf{r}) + \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{S}(\mathbf{x}(\mathbf{r}'))\, d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(\mathbf{r}, t) \quad \forall \mathbf{x} \in \mathcal{F}, \tag{3.2}$$

or to f_a defined by the right-hand side of (2.7):

$$f_a(t, \mathbf{x})(\mathbf{r}) = -\mathbf{L}\mathbf{x}(\mathbf{r}) + \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{x}(\mathbf{r}')\, d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(\mathbf{r}, t)\right) \quad \forall \mathbf{x} \in \mathcal{F}. \tag{3.3}$$

We have the following proposition.

PROPOSITION 3.2. *If the following two hypotheses are satisfied:*

1. *the connectivity function* $\mathbf{W}$ *is in* $C(\mathrm{J}; \mathbf{C}_{\mathrm{n}\times\mathrm{n}}(\Omega \times \Omega))$ *(see section* A.2*),*
2. *the external current* $\mathbf{I}_{\mathrm{ext}}$ *is in* $C(\mathrm{J}; \mathbf{C}_{\mathrm{n}}(\Omega))$*,*

then the mappings f_v *and* f_a *are from* $\mathrm{J}\times\mathcal{F}$ *to* $\mathcal{F}$*, continuous, and Lipschitz continuous with respect to their second argument, uniformly with respect to the first (*$\mathbf{C}_{n\times n}(\Omega\times\Omega)$ *and* $\mathbf{C}_n(\Omega)$ *are defined in section* A.1*).*

Proof. Let $t \in \mathrm{J}$ and $\mathbf{x} \in \mathcal{F}$. We introduce the mapping

$$F_v : (t, \mathbf{x}) \rightarrow F_v(t, \mathbf{x}) \quad \text{such that} \quad F_v(t, \mathbf{x})(\mathbf{r}) = \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{S}(\mathbf{x}(\mathbf{r}'))\, d\mathbf{r}'. \tag{3.4}$$

$F_v(t, \mathbf{x})$ is well defined for all $\mathbf{r} \in \Omega$ because, thanks to the first hypothesis, it is the integral of the continuous function $\mathbf{W}(\mathbf{r}, ., t)\mathbf{S}(\mathbf{x}(.))$ on a compact domain. For all $\mathbf{r}' \in \Omega$, $\mathbf{W}(., \mathbf{r}', t)\mathbf{S}(\mathbf{x}(\mathbf{r}'))$ is continuous (first hypothesis again) and we have (Lemma A.1)

$$\|\mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{S}(\mathbf{x}(\mathbf{r}'))\|_\infty \le \|\mathbf{W}(., ., t)\|_{n\times n,\infty} \|\mathbf{S}(\mathbf{x}(\mathbf{r}'))\|_\infty.$$

Since $\|\mathbf{S}(\mathbf{x}(.))\|_\infty$ is bounded, it is integrable in Ω, and we conclude that $F_v(t, \mathbf{x})$ is continuous on Ω. Then it is easy to see that $f_v(t, \mathbf{x})$ is well defined and belongs to $\mathcal{F}$.

Let us prove that f_v is continuous:

$$\begin{aligned} f_v(t, \mathbf{x}) - f_v(s, \mathbf{y}) = -\mathbf{L}(\mathbf{x} - \mathbf{y}) &+ \int_\Omega \left(\mathbf{W}(\cdot, \mathbf{r}', t)\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{W}(\cdot, \mathbf{r}', s)\mathbf{S}(\mathbf{y}(\mathbf{r}'))\right)\, d\mathbf{r}' \\ &+ \mathbf{I}_{\mathrm{ext}}(\cdot, t) - \mathbf{I}_{\mathrm{ext}}(\cdot, s) \\ = -\mathbf{L}(\mathbf{x} - \mathbf{y}) &+ \int_\Omega (\mathbf{W}(\cdot, \mathbf{r}', t) - \mathbf{W}(\cdot, \mathbf{r}', s))\mathbf{S}(\mathbf{x}(\mathbf{r}'))\, d\mathbf{r}' \\ &+ \int_\Omega \mathbf{W}(\cdot, \mathbf{r}', s)(\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{S}(\mathbf{y}(\mathbf{r}')))\, d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(\cdot, t) - \mathbf{I}_{\mathrm{ext}}(\cdot, s). \end{aligned}$$

It follows from Lemma 3.1 that

$$\begin{aligned} \|f_v(t, \mathbf{x}) - f_v(s, \mathbf{y})\|_{n,\infty} \le{} & \|\mathbf{L}\|_\infty \, \|\mathbf{x} - \mathbf{y}\|_{n,\infty} + |\Omega|\, S_m \|\mathbf{W}(\cdot, \cdot, t) - \mathbf{W}(\cdot, \cdot, s)\|_{n\times n,\infty} \\ & + |\Omega|\, \|\mathbf{W}(\cdot, \cdot, s)\|_{n\times n,\infty} DS_m \, \|\mathbf{x} - \mathbf{y}\|_{n,\infty} + \|\mathbf{I}_{\mathrm{ext}}(\cdot, t) - \mathbf{I}_{\mathrm{ext}}(\cdot, s)\|_{n,\infty}. \end{aligned}$$

Because of the hypotheses we can choose $|t - s|$ small enough so that $\|\mathbf{W}(\cdot, \cdot, t) - \mathbf{W}(\cdot, \cdot, s)\|_{n\times n,\infty}$ and $\|\mathbf{I}_{\mathrm{ext}}(\cdot, t) - \mathbf{I}_{\mathrm{ext}}(\cdot, s)\|_{n,\infty}$ are arbitrarily small. Similarly, since $\mathbf{W}$ is continuous on the compact interval J, it is bounded there and $\|\mathbf{W}(\cdot, \cdot, s)\|_{n\times n,\infty} \le w > 0$ for all $s \in \mathrm{J}$. This proves the continuity of f_v.

It follows from the previous inequality that

$$\|f_v(t,\mathbf{x}) - f_v(t,\mathbf{y})\|_{n,\infty} \le \|\mathbf{L}\|_\infty \, \|\mathbf{x}-\mathbf{y}\|_{n,\infty} + |\Omega| \, \|\mathbf{W}(\cdot,\cdot,t)\|_{n\times n,\infty} DS_m \, \|\mathbf{x}-\mathbf{y}\|_{n,\infty},$$

and because $\|\mathbf{W}(\cdot,\cdot,t)\|_{n\times n,\infty} \le w > 0$ for all ts in J, this proves the Lipschitz continuity of f_v with respect to its second argument, uniformly with respect to the first.

A very similar proof applies to f_a. □

We continue with the proof that there exists a unique solution to the abstract initial value problem (3.1) in the two cases of interest.

Proposition 3.3. *Subject to the hypotheses of Proposition* 3.2 *for any element* $\mathbf{V}_0$ *(resp.,* $\mathbf{A}_0$*) of* $\mathcal{F}$ *there is a unique solution* $\mathbf{V}$ *(resp.,* $\mathbf{A}$*), defined on a subinterval of* J *containing* 0 *and continuously differentiable, of the abstract initial value problem* (3.1) *for* $f = f_v$ *(resp.,* $f = f_a$*).*

Proof. All conditions of the Picard–Lindelöf theorem on differential equations in Banach spaces [9, 2] are satisfied; hence the proposition is proved. □

This solution, defined on the subinterval J of $\mathbb{R}$ can in fact be extended to the whole real line, and we have the following proposition.

Proposition 3.4. *If the following two hypotheses are satisfied:*

1. *the connectivity function* $\mathbf{W}$ *is in* $C(\mathbb{R};\mathbf{C}_{\mathrm{n\times n}}(\Omega\times\Omega))$*,*
2. *the external current* $\mathbf{I}_{\mathrm{ext}}$ *is in* $C(\mathbb{R};\mathbf{C}_{\mathrm{n}}(\Omega))$*,*

then for any function $\mathbf{V}_0$ *(resp.,* $\mathbf{A}_0$*) in* $\mathcal{F}$ *there is a unique solution* $\mathbf{V}$ *(resp.,* $\mathbf{A}$*), defined on* $\mathbb{R}$ *and continuously differentiable, of the abstract initial value problem* (3.1) *for* $f = f_v$ *(resp.,* $f = f_a$*).*

Proof. In Theorem B.1, we prove the existence of a constant $\tau > 0$ such that for any initial condition $(t_0, \mathbf{V}_0) \in \mathbb{R}\times\mathcal{F}$, there is a unique solution defined on the closed interval $[t_0-\tau,\, t_0+\tau]$. We can then cover the real line with such intervals and finally obtain the global existence and uniqueness of the solution of the initial value problem. □

3.2. Homogeneous solution. A homogeneous solution to (2.6) or (2.7) is a solution $\mathbf{U}$ that does not depend upon the space variable $\mathbf{r}$ for a given homogeneous input $\mathbf{I}_{\mathrm{ext}}(t)$ and a constant initial condition $\mathbf{U}_0$. If such a solution $\mathbf{U}(t)$ exists, then it satisfies

$$\mathbf{U}'(t) = -\mathbf{L}\mathbf{U}(t) + \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{S}(\mathbf{U}(t))\,d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(t)$$

in the case of (2.6) and

$$\mathbf{U}'(t) = -\mathbf{L}\mathbf{U}(t) + \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{U}(t)\,d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(t)\right)$$

in the case of (2.7). The integral $\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{S}(\mathbf{U}(t))\,d\mathbf{r}'$ is equal to $\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\,d\mathbf{r}'\right)$ $\cdot\mathbf{S}(\mathbf{U}(t))$. The integral $\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{U}(t)\,d\mathbf{r}'$ is equal to $\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\,d\mathbf{r}'\right)\mathbf{U}(t)$. They must be independent of the position $\mathbf{r}$. Hence a necessary condition for the existence of a homogeneous solution is that

$$\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\,d\mathbf{r}' = \overline{\mathbf{W}}(t), \tag{3.5}$$

where the $n\times n$ matrix $\overline{\mathbf{W}}(t)$ does not depend on the spatial coordinate $\mathbf{r}$.

In the special case where $\mathbf{W}(\mathbf{r}, \mathbf{r}', t)$ is translation invariant, $\mathbf{W}(\mathbf{r}, \mathbf{r}', t) \equiv \mathbf{W}(\mathbf{r} - \mathbf{r}', t)$, the condition is not satisfied in general because of the border of Ω. In all cases, the homogeneous solutions satisfy the differential equation

$$\mathbf{U}'(t) = -\mathbf{L}\mathbf{U}(t) + \overline{\mathbf{W}}(t)\mathbf{S}(\mathbf{U}(t)) + \mathbf{I}_{\text{ext}}(t) \tag{3.6}$$

for (2.6) and

$$\mathbf{U}'(t) = -\mathbf{L}\mathbf{U}(t) + \mathbf{S}\left(\overline{\mathbf{W}}(t)\mathbf{U}(t)) + \mathbf{I}_{\text{ext}}(t)\right) \tag{3.7}$$

for (2.7), with initial condition $\mathbf{U}(0) = \mathbf{U}_0$, a vector of $\mathbb{R}^n$. The following theorem gives a sufficient condition for the existence of a homogeneous solution.

Theorem 3.5. *If the external current $\mathbf{I}_{\text{ext}}(t)$ and the connectivity matrix $\overline{\mathbf{W}}(t)$ are continuous on some closed interval* J *containing* 0*, then for all vectors $\mathbf{U}_0$ of $\mathbb{R}^n$, there exists a unique solution $\mathbf{U}(t)$ of* (3.6) *or* (3.7) *defined on a subinterval* J_0 *of* J *containing* 0 *such that $\mathbf{U}(0) = \mathbf{U}_0$.*

Proof. The proof is an application of Cauchy's theorem on differential equations. Consider the mapping $f_{hv} : \mathbb{R}^n \times \mathrm{J} \to \mathbb{R}^n$ defined by

$$f_{hv}(\mathbf{x}, t) = -\mathbf{L}\mathbf{x} + \overline{\mathbf{W}}(t)\mathbf{S}(\mathbf{x}) + \mathbf{I}_{\text{ext}}(t).$$

We have

$$\|f_{hv}(\mathbf{x}, t) - f_{hv}(\mathbf{y}, t)\|_\infty \leq \|\mathbf{L}\|_\infty \|\mathbf{x} - \mathbf{y}\|_\infty + \|\overline{\mathbf{W}}(t)\|_\infty \|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{y})\|_\infty.$$

It follows from Lemma 3.1 that $\|\mathbf{S}(\mathbf{x}) - \mathbf{S}(\mathbf{y})\|_\infty \leq DS_m \|\mathbf{x} - \mathbf{y}\|_\infty$, and, since $\overline{\mathbf{W}}$ is continuous on the compact interval J, it is bounded there by $w > 0$ and

$$\|f_{hv}(\mathbf{x}, t) - f_{hv}(\mathbf{y}, t)\|_\infty \leq (\|\mathbf{L}\|_\infty + wDS_m)\|\mathbf{x} - \mathbf{y}\|_\infty$$

for all $\mathbf{x}, \mathbf{y}$ of $\mathbb{R}^n$ and all $t \in \mathrm{J}$. A similar proof applies to (3.7), and the conclusion of the proposition follows. □

As in Proposition 3.4, this existence and uniqueness result extends to the whole time real line if $\mathbf{I}$ and $\overline{\mathbf{W}}$ are continuous on $\mathbb{R}$.

This homogeneous solution can be seen as describing a state where the columns of the continuum are synchronized: they receive the same input $\mathbf{I}_{\text{ext}}(t)$ and produce the same output $\mathbf{U}(t)$.

3.3. Some remarks about the case $\boldsymbol{\Omega} = \mathbb{R}^q$. A significant amount of work has been done on equations of the type (2.6) or (2.7) in the case of a one-dimensional infinite continuum, $\Omega = \mathbb{R}$, or a two-dimensional infinite continuum, $\Omega = \mathbb{R}^2$. The reader is referred to the review papers by Ermentrout [10] and by Coombes [6] as well as to [33, 14, 35].

Aside from the fact that an infinite cortex is unrealistic, the case $\Omega = \mathbb{R}^q$ raises some mathematical questions. Indeed, the choice of the functional space $\mathcal{F}$ is problematic. A natural idea would be to choose $\mathcal{F} = \mathbf{L}^2_n(\mathbb{R}^q)$, the space of square-integrable functions with values in $\mathbb{R}^n$; see section A.1. If we make this choice, we immediately encounter the problem that the homogeneous solutions (constant with respect to the space variable) do not belong to that space. A further difficulty is that $\mathbf{S}(\mathbf{x})$ does not in general belong to $\mathcal{F}$ if $\mathbf{x}$ does. As shown in this article, these difficulties vanish if Ω is compact.

4. Absolute stability of the general solution. We investigate the absolute stability of a solution to (2.6) and (2.7) for a given input $\mathbf{I}_{\text{ext}}$. Proposition 3.4 guarantees that for a given initial condition there exists a unique solution to (2.6) or (2.7) defined for all times.

In order to investigate its absolute stability we choose a different initial condition, which is a way to perturb the solution (in effect the only way because of the existence uniqueness proposition, Proposition 3.4), and look for sufficient conditions for the new solution to converge toward the original one. Absolute stability implies linear stability, which is studied by perturbing the solution by adding to it a small function and performing a first-order Taylor expansion of the equations, thereby obtaining a perturbed equation. One then usually has to make some assumptions about the spatio-temporal form of the perturbation, e.g., that it is separable in time and space, ending up with a nontrivial eigenvalue problem which has to be solved in order to find sufficient conditions for the perturbation to converge to 0, up to first-order [6, 10, 12, 14, 28, 33, 34, 24, 35].

This is also the case of [1] and [7]; those authors study the convolution case for $n = q = 1$ but incorporate propagation delays. Linear stability is local because it is derived for a particular solution. The functional analysis approach that we use in this paper allows us to find simple sufficient conditions for the absolute stability of the system, and hence for all its solutions, regardless of the initial condition or input. In this sense it is a global approach. This is achieved by constructing a Lyapunov function measuring some distance between two state vectors at each time instant. This function has a single minimum corresponding to the equality of the states. One then finds sufficient conditions for the time derivative of this function to be strictly negative, thereby guaranteeing the asymptotic equality of the states. This approach has been followed by fewer people. In [23] the authors study the case where $\mathbf{W}(\mathbf{r}, \mathbf{r}')$ is symmetric with respect to the space variables $\mathbf{r}$ and $\mathbf{r}'$ for $n = q = 1$ for a finite interval and add the translation invariance assumption when the interval is infinite. They do not study the case of general time-varying input currents.

Absolute stability is a relevant concept for systems of neurons. Indeed, absolutely stable systems forget their initial state exponentially fast but do not forget their input. Hence such systems can differentiate distinct stimuli by converging to the corresponding states without being influenced by their initial state. This property is desirable, for example, in modeling visual perception: different forms elicit different percepts, but the percepts should not depend on the initial state of the visual system. We first look at the general case and then at the convolution case.

4.1. The general case. We define a number of matrices and linear operators that are useful in what follows.

Definition 4.1. *Let*

$$\mathbf{W}_{cm} = \mathbf{W} D\mathbf{S}_m, \quad \mathbf{W}_{mc} = D\mathbf{S}_m\mathbf{W}.$$

Consider also the linear operators, noted g, g_m, and h_m, defined on $\mathcal{F}$:

$$g(\mathbf{x})(\mathbf{r}, t) = \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{x}(\mathbf{r}')\, d\mathbf{r}' \quad \forall \mathbf{x} \in \mathcal{F},$$

$$g_m(\mathbf{x})(\mathbf{r}, t) = \int_\Omega \mathbf{W}_{cm}(\mathbf{r}, \mathbf{r}', t)\mathbf{x}(\mathbf{r}')\, d\mathbf{r}' \quad \forall \mathbf{x} \in \mathcal{F},$$

and

$$h_m(\mathbf{x})(\mathbf{r},t) = \int_\Omega \mathbf{W}_{mc}(\mathbf{r},\mathbf{r}',t)\mathbf{x}(\mathbf{r}')\,d\mathbf{r}' \quad \forall \mathbf{x} \in \mathcal{F}.$$

We start with a lemma.

LEMMA 4.2. *With the hypotheses of Proposition* 3.2, *the operators* g, g_m, *and* h_m *are compact operators from* $\mathcal{F}$ *to* $\mathcal{F}$ *for each time* $t \in \mathrm{J}$.

Proof. This is a direct application of the theory of Fredholm's integral equations [9]. We prove it for g.

Because of hypothesis 1 in Proposition 3.2, at each time instant t in J, $\mathbf{W}$ is continuous on the compact set $\Omega \times \Omega$; therefore, it is uniformly continuous. Hence, for each $\varepsilon > 0$ there exists $\eta(t) > 0$ such that $\|\mathbf{r}_1 - \mathbf{r}_2\| \le \eta(t)$ implies that $\|\mathbf{W}(\mathbf{r}_1,\mathbf{r}',t) - \mathbf{W}(\mathbf{r}_2,\mathbf{r}',t)\|_\infty \le \varepsilon$ for all $\mathbf{r}' \in \Omega$, and, for all $\mathbf{x} \in \mathcal{F}$,

$$\|g(\mathbf{x})(\mathbf{r}_1,t) - g(\mathbf{x})(\mathbf{r}_2,t)\|_\infty \le \varepsilon|\Omega|\|\mathbf{x}\|_{n,\infty}.$$

This shows that the image $g(B)$ of any bounded subset B of $\mathcal{F}$ is equicontinuous.

Similarly, if we set $w(t) = \|\mathbf{W}(.,.,t)\|_{n\times n,\infty}$, we have $\|g(\mathbf{x})(\mathbf{r},t)\|_\infty \le w(t)|\Omega|\|\mathbf{x}\|_{n,\infty}$. This shows that for every $\mathbf{r} \in \Omega$, the set $\{\mathbf{y}(\mathbf{r}),\, \mathbf{y} \in g(B)\}$ is bounded in $\mathbb{R}^n$ and hence relatively compact. From the Arzelà–Ascoli theorem, we conclude that the subset $g(B)$ of $\mathcal{F}$ is relatively compact for all $t \in \mathrm{J}$. And so the operator is compact.

The same proof applies to g_m and h_m. □

To study the absolute stability of the solutions of (2.6) and (2.7), it is convenient to use an inner product on $\mathcal{F}$. It turns out that the natural inner product will pave the way for the generalization in section 6. We therefore consider the pre-Hilbert space $\mathcal{G}$ defined on $\mathcal{F}$ by the usual inner product

$$\langle \mathbf{x},\, \mathbf{y}\rangle = \int_\Omega \mathbf{x}(\mathbf{r})^T\mathbf{y}(\mathbf{r})\,d\mathbf{r}.$$

We denote $\|\mathbf{x}\|_{n,2}$ as the corresponding norm to distinguish it from $\|\mathbf{x}\|_{n,\infty}$; see section A.1. It is easy to show that all previously defined operators are also compact operators from $\mathcal{G}$ to $\mathcal{G}$. We have the following lemma.

LEMMA 4.3. g, g_m, *and* h_m *are compact operators from* $\mathcal{G}$ *to* $\mathcal{G}$ *for each time* $t \in \mathrm{J}$.

Proof. We give the proof for g.

The identity mapping $\mathbf{x} \to \mathbf{x}$ from $\mathcal{F}$ to $\mathcal{G}$ is continuous since $\|\mathbf{x}\|_{n,2} \le \sqrt{n|\Omega|}\,\|\mathbf{x}\|_{n,\infty}$. Consider now g as a mapping from $\mathcal{G}$ to $\mathcal{F}$. As in the proof of Lemma 4.2, for each $\varepsilon > 0$ there exists $\eta(t) > 0$ such that $\|\mathbf{r}_1 - \mathbf{r}_2\| \le \eta(t)$ implies $\|\mathbf{W}(\mathbf{r}_1,\mathbf{r}',t) - \mathbf{W}(\mathbf{r}_2,\mathbf{r}',t)\|_\infty \le \varepsilon$ for all $\mathbf{r}' \in \Omega$. Therefore, the ith coordinate $g^i(\mathbf{x})(\mathbf{r}_1,t) - g^i(\mathbf{x})(\mathbf{r}_2,t)$ satisfies (Cauchy–Schwarz inequalities)

$$\begin{aligned}|g^i(\mathbf{x})(\mathbf{r}_1,t) - g^i(\mathbf{x})(\mathbf{r}_2,t)| &\le \sum_j \int_\Omega |W_{ij}(\mathbf{r}_1,\mathbf{r}',t) - W_{ij}(\mathbf{r}_2,\mathbf{r}',t)|\,|x_j(\mathbf{r}')|\,d\mathbf{r}' \\ &\le \varepsilon \sum_j \int_\Omega |x_j(\mathbf{r}')|\,d\mathbf{r}' \le \varepsilon\sqrt{|\Omega|}\sum_j \left(\int_\Omega |x_j(\mathbf{r}')|^2\,d\mathbf{r}'\right)^{1/2} \le \varepsilon\sqrt{n\,|\Omega|}\|\mathbf{x}\|_{n,2},\end{aligned}$$

and the image $g(B)$ of any bounded set B of $\mathcal{G}$ is equicontinuous. Similarly, if we set $w(t) = \|\mathbf{W}(.,.,t)\|_{n\times n,\infty}$ in $\Omega \times \Omega$, we have $|g^i(\mathbf{x})(\mathbf{r},t)| \le w(t)\sqrt{n\,|\Omega|}\,\|\mathbf{x}\|_{n,2}$. The same reasoning as in Lemma 4.2 shows that the operator $\mathbf{x} \to g(\mathbf{x})$ from $\mathcal{G}$ to $\mathcal{F}$ is compact, and since the identity from $\mathcal{F}$ to $\mathcal{G}$ is continuous, $\mathbf{x} \to g(\mathbf{x})$ is compact from $\mathcal{G}$ to $\mathcal{G}$.

The same proof applies to g_m and h_m. □

We proceed with the following lemma.

LEMMA 4.4. *The adjoint g^* of the operator g of $\mathcal{G}$ is the operator defined by*

$$g^*(\mathbf{x})(\mathbf{r}, t) = \int_\Omega \mathbf{W}^T(\mathbf{r}', \mathbf{r}, t)\mathbf{x}(\mathbf{r}')\, d\mathbf{r}'.$$

It is a compact operator. Similar results apply to g_m^* and h_m^*.

Proof. The adjoint, if it exists, is defined by the condition $\langle g(\mathbf{x}), \mathbf{y}\rangle = \langle \mathbf{x}, g^*(\mathbf{y})\rangle$ for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{G}$. We have

$$\begin{aligned}\langle g(\mathbf{x}),\, \mathbf{y}\rangle &= \int_\Omega \mathbf{y}(\mathbf{r})^T \left(\int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{x}(\mathbf{r}')\, d\mathbf{r}'\right) d\mathbf{r} \\ &= \int_\Omega \mathbf{x}(\mathbf{r}')^T \left(\int_\Omega \mathbf{W}^T(\mathbf{r}, \mathbf{r}', t)\mathbf{y}(\mathbf{r})\, d\mathbf{r}\right) d\mathbf{r}',\end{aligned}$$

from which the conclusion follows. Since $\mathcal{G}$ is not a Hilbert space, the adjoint of a compact operator is not necessarily compact. But the proof of compactness of g in Lemma 4.3 extends easily to g^*. □

We finally prove two useful lemmas that will complete our toolbox for the proof of the main results of this section.

LEMMA 4.5. *Given a diagonal matrix $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$, with $d_1, \ldots, d_n \in \mathbf{L}^\infty(\Omega)$ and a function $\mathbf{x} \in \mathcal{G}$, we have*

$$\|\mathbf{D}\mathbf{x}\|_{n,2} \le \max_i(\|d_i\|_\infty)\|\mathbf{x}\|_{n,2}.$$

Proof.

$$\|\mathbf{D}\mathbf{x}\|_{n,2}^2 = \int_\Omega \mathbf{x}(\mathbf{r})^T\mathbf{D}^2(\mathbf{r})\mathbf{x}(\mathbf{r})\, d\mathbf{r} = \sum_i \int_\Omega d_i^2(\mathbf{r})\mathbf{x}_i^2(\mathbf{r})\, d\mathbf{r} \le \sum_i \|d_i\|_\infty^2 \int_\Omega \mathbf{x}_i^2(\mathbf{r})\, d\mathbf{r},$$

from which the result follows. □

LEMMA 4.6. *$\|g\|_\mathcal{G}$, $\|g_m\|_\mathcal{G}$, and $\|h_m\|_\mathcal{G}$ satisfy the inequalities*

$$\|g_m\|_\mathcal{G} \le DS_m\, \|g\|_\mathcal{G} \quad \textit{and} \quad \|h_m\|_\mathcal{G} \le DS_m\, \|g\|_\mathcal{G}\,,$$

where DS_m is defined in Definition 2.1.

Proof. By definition

$$\|g_m\|_\mathcal{G} = \sup_{\|\mathbf{x}\|_{n,2}\le 1} \frac{\|g_m(\mathbf{x})\|_{n,2}}{\|\mathbf{x}\|_{n,2}} = \sup_{\|\mathbf{x}\|_{n,2}\le 1} \frac{\|g(D\mathbf{S}_m\mathbf{x})\|_{n,2}}{\|\mathbf{x}\|_{n,2}}.$$

Let $\mathbf{y} = D\mathbf{S}_m\mathbf{x}$. Since $\{\mathbf{x} \in \mathcal{G},\ \|\mathbf{x}\|_{n,2} \le 1\} \subset \{\mathbf{x} \in \mathcal{G},\ \|D\mathbf{S}_m\mathbf{x}\|_{n,2} \le DS_m\}$ (Lemma 4.5),

$$\begin{aligned}\|g_m\|_\mathcal{G} &\le \sup_{\|\mathbf{y}\|_{n,2}\le DS_m} \frac{\|g(\mathbf{y})\|_{n,2}}{\|D\mathbf{S}_m^{-1}\mathbf{y}\|_{n,2}} = \sup_{\|\mathbf{y}\|_{n,2}\le 1} \frac{\|g(\mathbf{y})\|_{n,2}}{\|D\mathbf{S}_m^{-1}\mathbf{y}\|_{n,2}} \\ &\le \sup_{\|\mathbf{y}\|_{n,2}\le 1} \frac{\|g(\mathbf{y})\|_{n,2}}{\|\mathbf{y}\|_{n,2}} \cdot \sup_{\|\mathbf{y}\|_{n,2}\le 1} \frac{\|\mathbf{y}\|_{n,2}}{\|D\mathbf{S}_m^{-1}\mathbf{y}\|_{n,2}} \le \|g\|_\mathcal{G}\, DS_m.\end{aligned}$$

The last inequality is also obtained from Lemma 4.5, which is used again to prove the inequality for h_m: $h_m = D\mathbf{S}_m g$ and $\|D\mathbf{S}_m g(\mathbf{x})\|_{n,2} \leq DS_m \|g(\mathbf{x})\|_{n,2}$ for all $\mathbf{x} \in \mathcal{G}$, from which the result follows. □

We show in Appendix D a table summarizing the main notation introduced so far for future reference.

We now state an important result of this section.

THEOREM 4.7. *A sufficient condition for the absolute stability of a solution to* (2.6) *is*

$$DS_m \tau_{\max} \|g\|_{\mathcal{G}} < 1, \tag{4.1}$$

where $\|.\|_{\mathcal{G}}$ *is the operator norm.*

Proof. Let us denote $\underline{\mathbf{S}}$ as the function $D\mathbf{S}_m^{-1}\mathbf{S}$ and rewrite (2.6) as

$$\mathbf{V}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{V}(\mathbf{r},t) + \int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)\,\underline{\mathbf{S}}(\mathbf{V}(\mathbf{r}',t))\,d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t).$$

Let $\mathbf{U}$ be its unique solution with initial conditions $\mathbf{U}(0) = \mathbf{U}_0$, an element of $\mathcal{G}$. Let also $\mathbf{V}$ be the unique solution of the same equation with different initial conditions $\mathbf{V}(0) = \mathbf{V}_0$, another element of $\mathcal{G}$. We introduce the new function $\mathbf{X} = \mathbf{V} - \mathbf{U}$ which satisfies

$$\begin{aligned}\mathbf{X}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{X}(\mathbf{r},t) + \int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)\,\mathbf{H}(\mathbf{X},\mathbf{U})(\mathbf{r}',t)\,d\mathbf{r}'\\ = -\mathbf{L}\mathbf{X}(\mathbf{r},t) + g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))(\mathbf{r},t),\end{aligned} \tag{4.2}$$

where the vector $\mathbf{H}(\mathbf{X},\mathbf{U})$ is given by $\mathbf{H}(\mathbf{X},\mathbf{U})(\mathbf{r},t)) = \underline{\mathbf{S}}(\mathbf{V}(\mathbf{r},t)) - \underline{\mathbf{S}}(\mathbf{U}(\mathbf{r},t)) = \underline{\mathbf{S}}(\mathbf{X}(\mathbf{r},t) + \mathbf{U}(\mathbf{r},t)) - \underline{\mathbf{S}}(\mathbf{U}(\mathbf{r},t))$. Consider now the functional (Lyapunov function)

$$V(\mathbf{X}) = \frac{1}{2}\left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathbf{X} \right\rangle,$$

where the symmetric positive definite matrix $\mathbf{L}$ can be seen as defining a metric on the state space. Its time derivative is $\left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathbf{X}_t \right\rangle$. We replace $\mathbf{X}_t$ by its value from (4.2) in this expression to obtain

$$\frac{dV(\mathbf{X})}{dt} = -\left\langle \mathbf{X},\, \mathbf{X} \right\rangle + \left\langle \mathbf{X},\, \mathbf{L}^{-1} g_m(\mathbf{H}(\mathbf{X},\mathbf{U})) \right\rangle.$$

We consider the second term on the right-hand side of this equation:

$$\begin{aligned}\left| \left\langle \mathbf{X},\, \mathbf{L}^{-1} g_m(\mathbf{H}(\mathbf{X},\mathbf{U})) \right\rangle \right| &\leq \|\mathbf{X}\|_{n,2}\, \|\mathbf{L}^{-1} g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\|_{n,2}\\ &\leq \tau_{\max}\|\mathbf{X}\|_{n,2}\, \|g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\|_{n,2} \leq \tau_{\max}\|\mathbf{X}\|_{n,2}\, \|g_m\|_{\mathcal{G}} \|\mathbf{H}(\mathbf{X},\mathbf{U})\|_{n,2}.\end{aligned} \tag{4.3}$$

Using a zeroth-order Taylor expansion with integral remainder, as in the proof of Lemma 3.1, we write $\mathbf{H}(\mathbf{X},\mathbf{U}) = \mathcal{D}_m\mathbf{X}$, where $\mathcal{D}_m$ is a diagonal matrix whose diagonal elements are continuous functions with values between 0 and 1:

$$\mathcal{D}_m(\mathbf{r},t) = \int_0^1 D\underline{\mathbf{S}}(\mathbf{U}(\mathbf{r},t) + \zeta(\mathbf{V}(\mathbf{r},t) - \mathbf{U}(\mathbf{r},t)))\,d\zeta.$$

Hence, according to Lemma 4.5,

$$\|\mathbf{H}(\mathbf{X},\mathbf{U})\|_{n,2} = \|\mathcal{D}_m\mathbf{X}\|_{n,2} \leq \|\mathbf{X}\|_{n,2}.$$

We use this result and Lemma 4.6 in (4.3) to obtain

$$| \left\langle \mathbf{X},\, \mathbf{L}^{-1} g_m(\mathbf{H}(\mathbf{X},\mathbf{U})) \right\rangle | \leq \tau_{\max} DS_m \, \|g\|_{\mathcal{G}} \, \|\mathbf{X}\|_{n,2}^2,$$

and the conclusion follows. □

An identical sufficient condition holds for the stability of a solution to (2.7).

Theorem 4.8. *A sufficient condition for the absolute stability of a solution to* (2.7) *is*

$$DS_m \tau_{\max} \, \|g\|_{\mathcal{G}} < 1.$$

Proof. Let $\mathbf{U}$ be the unique solution of (2.7) with an external current $\mathbf{I}_{\text{ext}}(\mathbf{r},t)$ and initial conditions $\mathbf{U}(0) = \mathbf{U}_0$. As in the proof of Theorem 4.7, we introduce the new function $\mathbf{X} = \mathbf{V} - \mathbf{U}$, where $\mathbf{V}$ is the unique solution of the same equation with different initial conditions. We have

$$\begin{aligned} \mathbf{X}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{X}(\mathbf{r},t) &+ \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{V}(\mathbf{r}',t)\, d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t) \right) \\ &- \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{U}(\mathbf{r}',t)\, d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t) \right). \end{aligned} \tag{4.4}$$

Using a zeroth-order Taylor expansion, as in the proof of Lemma 3.1, this equation can be rewritten as

$$\begin{aligned} \mathbf{X}_t(\mathbf{r},t) = -\mathbf{L}\mathbf{X}(\mathbf{r},t) + \Bigg(\int_0^1 D\mathbf{S}\Bigg(&\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{U}(\mathbf{r}'t)\, d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t) \\ &+ \zeta \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{X}(\mathbf{r}',t)\, d\mathbf{r}' \Bigg) d\zeta \Bigg) \left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{X}(\mathbf{r}',t)\, d\mathbf{r}' \right). \end{aligned}$$

We use the same functional as in the proof of Theorem 4.7:

$$V(\mathbf{X}) = \frac{1}{2} \left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathbf{X} \right\rangle.$$

Its time derivative is readily obtained with the help of (4.4),

$$\frac{dV(\mathbf{X})}{dt} = -\left\langle \mathbf{X},\, \mathbf{X} \right\rangle + \left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathcal{D}_m h_m(\mathbf{X}) \right\rangle, \tag{4.5}$$

where $\mathcal{D}_m$ is defined by

$$\begin{aligned} &\mathcal{D}_m(\mathbf{U},\mathbf{X},\mathbf{r},t) \\ &= \int_0^1 D\mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{U}(t)\, d\mathbf{r}' + \mathbf{I}_{\text{ext}}(\mathbf{r},t) + \zeta \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\mathbf{X}(\mathbf{r}',t)\, d\mathbf{r}' \right) D\mathbf{S}_m^{-1}\, d\zeta, \end{aligned}$$

a diagonal matrix whose diagonal elements are continuous functions with values between 0 and 1. We consider the second term on the right-hand side of (4.5) and use the property of matrix $\mathcal{D}_m$ and Lemma 4.6 to obtain

$$\begin{aligned} | \left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathcal{D}_m h_m(\mathbf{X}) \right\rangle | &\leq \|\mathbf{X}\|_{n,2} \|\mathbf{L}^{-1}\mathcal{D}_m h_m(\mathbf{X})\|_{n,2} \\ &\leq \tau_{\max} \|\mathbf{X}\|_{n,2} \, \|h_m(\mathbf{X})\|_{n,2} \leq \tau_{\max}\, DS_m \, \|g\|_{\mathcal{G}} \, \|\mathbf{X}\|_{n,2}^2, \end{aligned}$$

from which the result follows. □

Note that $\|g\|_{\mathcal{G}} = \|g\|_{\mathbf{L}^2}$ by density of $\mathcal{G}$ in $\mathbf{L}^2$ (see section 6). In Appendix A, we show how to compute such operator norms.

4.2. The convolution case. In the case where $\mathbf{W}$ is translation invariant, we can obtain a slightly easier to exploit sufficient condition for the stability of the solutions than in Theorems 4.7 and 4.8. We first consider the case of a general compact Ω and then the case where Ω is an interval. Translation invariance means that $\mathbf{W}(\mathbf{r}+\mathbf{a},\ \mathbf{r}'+\mathbf{a},\ t) = \mathbf{W}(\mathbf{r},\mathbf{r}',t)$ for all $\mathbf{a}$ such that $\mathbf{a}+\mathbf{r} \in \Omega$ and $\mathbf{a}+\mathbf{r}' \in \Omega$, so we can write $\mathbf{W}(\mathbf{r},\mathbf{r}',t) = \mathbf{W}(\mathbf{r}-\mathbf{r}',t)$. Hence $\mathbf{W}(\mathbf{r},t)$ must be defined for all $\mathbf{r} \in \widehat{\Omega} = \{\mathbf{r}-\mathbf{r}', \text{ with } \mathbf{r},\mathbf{r}' \in \Omega\}$, and we suppose it to be continuous on $\widehat{\Omega}$ for each t. $\widehat{\Omega}$ is a symmetric with respect to the origin, compact subset of $\mathbb{R}^q$.

4.2.1. General $\boldsymbol{\Omega}$. We denote $\mathbf{1}_A$ as the characteristic function of the subset A of $\mathbb{R}^q$ and $\mathbf{M}^* = \overline{\mathbf{M}}^T$ as the conjugate transpose of the complex matrix $\mathbf{M}$.

We prove the following theorem.

THEOREM 4.9. *If the eigenvalues of the Hermitian matrix*

$$\widetilde{\mathbf{W}}^*(\mathbf{f},t)\,\widetilde{\mathbf{W}}(\mathbf{f},t) \tag{4.6}$$

are strictly less than $(\tau_{\max} DS_m)^{-2}$ for almost all $\mathbf{f} \in \mathbb{R}^q$ and $t \in \mathrm{J}$, then the system (2.6) *is absolutely stable.*[1] *$\widetilde{\mathbf{W}}(\mathbf{f},t)$ is the Fourier transform with respect to the space variable $\mathbf{r}$ of $\mathbf{1}_{\widehat{\Omega}}(\mathbf{r})\,\mathbf{W}(\mathbf{r},t)$,*

$$\widetilde{\mathbf{W}}(\mathbf{f},t) = \int_{\widehat{\Omega}} \mathbf{W}(\mathbf{r},t) e^{-2i\pi \mathbf{r}\cdot\mathbf{f}}\, d\mathbf{r}.$$

Proof. We recall that

$$\|g\|_{\mathcal{G}}^2 = \sup_{\|\mathbf{x}\|_{n,2} \le 1} \frac{\|g(\mathbf{x})\|_{n,2}^2}{\|\mathbf{x}\|_{n,2}^2}.$$

We then note that, by definition,

$$\|g(\mathbf{x})\|_{n,2} = \|(\mathbf{1}_{\widehat{\Omega}}\mathbf{W}) \otimes (\mathbf{1}_\Omega \mathbf{x})\|_{\mathbb{R}^q,\,n,\,2},$$

where $\otimes$ indicates the convolution over $\mathbb{R}^q$. Parseval's theorem gives

$$\|(\mathbf{1}_{\widehat{\Omega}}\mathbf{W}) \otimes (\mathbf{1}_\Omega \mathbf{x})\|_{\mathbb{R}^q,\,n,\,2}^2 = \int_{\mathbb{R}^q} \widetilde{\mathbf{x}}^*(\mathbf{f},t)\widetilde{\mathbf{W}}^*(\mathbf{f},t)\widetilde{\mathbf{W}}(\mathbf{f},t)\widetilde{\mathbf{x}}(\mathbf{f},t)\, d\mathbf{f},$$

where $\widetilde{\mathbf{x}}$ is the Fourier transform of $\mathbf{1}_\Omega \mathbf{x}$.

As a Hermitian matrix, $\widetilde{\mathbf{W}}^*(\mathbf{f},t)\widetilde{\mathbf{W}}(\mathbf{f},t)$ can be rewritten as $\mathbf{U}^*(\mathbf{f},t)\mathbf{D}(\mathbf{f},t)\mathbf{U}(\mathbf{f},t)$, with $\mathbf{U}^*\mathbf{U} = \mathrm{Id_n}$ and $\mathbf{D}$ real and diagonal. In particular, $\mathbf{U}$ preserves length ($\|\mathbf{U}\mathbf{v}\|_2 = \|\mathbf{v}\|_2$). Besides, $\widetilde{\mathbf{W}}^*\widetilde{\mathbf{W}}$ is positive because for any complex vector $\mathbf{v}$,

$$\mathbf{v}^*\widetilde{\mathbf{W}}^*\widetilde{\mathbf{W}}\mathbf{v} = \|\widetilde{\mathbf{W}}\mathbf{v}\|_2^2 \ge 0.$$

So, all values of $\mathbf{D}$ are positive, and if the hypothesis of the theorem is satisfied, Lemma 4.5 yields

$$\begin{aligned}&\int_{\mathbb{R}^q} \widetilde{\mathbf{x}}^*(\mathbf{f},t)\widetilde{\mathbf{W}}^*(\mathbf{f},t)\widetilde{\mathbf{W}}(\mathbf{f},t)\widetilde{\mathbf{x}}(\mathbf{f},t)\, d\mathbf{f} = \|\sqrt{\mathbf{D}}\mathbf{U}\widetilde{\mathbf{x}}\|_{\mathbb{R}^q,\,n,\,2}^2 \\ &\quad\le (\tau_{\max} DS_m)^{-2}\|\mathbf{U}\widetilde{\mathbf{x}}\|_{\mathbb{R}^q,\,n,\,2}^2 = (\tau_{\max} DS_m)^{-2}\|\widetilde{\mathbf{x}}\|_{\mathbb{R}^q,\,n,\,2}^2 = (\tau_{\max} DS_m)^{-2}\|\mathbf{x}\|_{n,2}^2;\end{aligned}$$

[1] Note that since $\widetilde{\mathbf{W}}$ is continuous with respect to $\mathbf{f}$, some eigenvalues of the Hermitian matrix may be equal to $(\tau_{\max} DS_m)^{-2}$ on a domain of measure 0 of the $\mathbf{f}$-plane.

hence $\|g\|_{\mathcal{G}} < (\tau_{\max} D S_m)^{-1}$, and Theorem 4.7 applies. □

Since the sufficient condition for the absolute stability of the solution of the activation-based model is identical, we have the following theorem.

THEOREM 4.10. *If the eigenvalues of the Hermitian matrix*

$$\widetilde{\mathbf{W}}^*(\mathbf{f},t)\,\widetilde{\mathbf{W}}(\mathbf{f},t)$$

are strictly less than $(\tau_{\max} D S_m)^{-2}$ *for almost all* $\mathbf{f}$ *and* $t \in \mathrm{J}$, *then the system* (2.7) *is absolutely stable.* $\widetilde{\mathbf{W}}(\mathbf{f},t)$ *is the Fourier transform of* $\mathbf{1}_{\widehat{\Omega}}(\mathbf{r})\mathbf{W}(\mathbf{r},t)$ *with respect to the space variable* $\mathbf{r}$.

These two theorems are somewhat unsatisfactory since they replace a condition that must be satisfied over a countable set, the spectrum of a compact operator, as in Theorems 4.7 and 4.8, by a condition that must be satisfied over a continuum, i.e., $\mathbb{R}^q$. Nonetheless, one may consider that the computation of the Fourier transforms of the matrix $\mathbf{W}$, extended by zeros outside $\widehat{\Omega}$, is easier than that of the spectrum of the operator g, for which a method is given in section A.3.

4.2.2. Ω is an interval. In the case where Ω is an interval, i.e., an interval of $\mathbb{R}$ ($q = 1$), a parallelogram ($q = 2$), or a parallelepiped ($q = 3$), we can state different sufficient conditions. We can always assume that Ω is the q-dimensional interval $[0,1]^q$ by applying an affine change of coordinates. The connectivity matrix $\mathbf{W}$ is defined on $\mathrm{J} \times [-1,1]^{\mathrm{q}}$ and extended to a q-periodic function of periods 2 on $\mathrm{J} \times \mathbb{R}^{\mathrm{q}}$, reflecting periodic boundary conditions. Similarly, the state vectors $\mathbf{V}$ and $\mathbf{A}$ as well as the external current $\mathbf{I}_{\mathrm{ext}}$ defined on $\mathrm{J} \times [0,1]^{\mathrm{q}}$ are extended to q-periodic functions of the same periods over $\mathrm{J} \times \mathbb{R}^{\mathrm{q}}$ by padding them with zeros in the complement in the interval $[-1,1]^q$ of the interval $[0,1]^q$. $\mathcal{G}$ is now the space $\mathbf{L}^2_n(2)$ of the square-integrable q-periodic functions of periods 2 with values in $\mathbb{R}^n$.

We define the functions $\psi_{\mathbf{m}}(\mathbf{r}) \equiv e^{-\pi i(r_1 m_1 + \cdots + r_q m_q)}$ for $\mathbf{m} \in \mathbb{Z}^q$ and consider the matrix $\widetilde{\mathbf{W}}(\mathbf{m},t)$ whose elements are given by

$$\widetilde{W}_{ij}(\mathbf{m},t) = \int_{[0,2]^q} W_{ij}(\mathbf{r},t)\psi_{\mathbf{m}}(\mathbf{r})\,d\mathbf{r}, \quad 1 \le i,\, j \le n.$$

We recall the following definition.

DEFINITION 4.11. *The matrix* $\widetilde{\mathbf{W}}(\mathbf{m})$ *is the* $\mathbf{m}$*th element of the Fourier series of the periodic matrix function* $\mathbf{W}(\mathbf{r})$.

Theorems 4.9 and 4.10 can be stated in this framework.

THEOREM 4.12. *If the eigenvalues of the Hermitian matrix*

$$\widetilde{\mathbf{W}}^*(\mathbf{m},t)\,\widetilde{\mathbf{W}}(\mathbf{m},t) \tag{4.7}$$

are strictly less than $(\tau_{\max} D S_m)^{-2}$ *for all* $\mathbf{m} \in \mathbb{Z}^q$ *and all* $t \in \mathrm{J}$, *then the system* (2.6) *(resp.,* (2.7)*) is absolutely stable.* $\widetilde{\mathbf{W}}(\mathbf{m},t)$ *is the* $\mathbf{m}$*th element of the Fourier series of the q-periodic matrix function* $\mathbf{W}(\mathbf{r},t)$ *with periods* 2 *at time* t.

5. Absolute stability of the homogeneous solution.

We next investigate the absolute stability of a homogeneous solution to (2.6) and (2.7). As in the previous section, we distinguish the general and convolution cases.

5.1. The general case. The homogeneous solutions are characterized by the fact that they are spatially constant at each time instant. We consider the subspace $\mathcal{G}_c$ of $\mathcal{G}$ of the constant functions. We have the following lemma.

LEMMA 5.1. *$\mathcal{G}_c$ is a complete linear subspace of $\mathcal{G}$. The orthogonal projection operator $\mathcal{P}_{\mathcal{G}_c}$ from $\mathcal{G}$ to $\mathcal{G}_c$ is defined by*

$$\mathcal{P}_{\mathcal{G}_c}(\mathbf{x}) = \overline{\mathbf{x}} = \frac{1}{|\Omega|}\int_\Omega \mathbf{x}(\mathbf{r})\,d\mathbf{r}.$$

The orthogonal complement $\mathcal{G}_c^\perp$ of $\mathcal{G}_c$ is the subset of functions of $\mathcal{G}$ that have a zero average. The orthogonal projection[2] *operator $\mathcal{P}_{\mathcal{G}_c^\perp}$ is equal to $\mathrm{Id} - \mathcal{P}_{\mathcal{G}_c}$. We also have*

$$\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{M}\mathbf{x} = \mathbf{M}\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{x} \quad \forall\,\mathbf{x}\in\mathcal{G},\,\mathbf{M}\in\mathcal{M}_{n\times n}. \tag{5.1}$$

Proof. The constant functions are clearly in $\mathcal{G}$. Any Cauchy sequence of constants is converging to a constant; hence $\mathcal{G}_c$ is closed in the pre-Hilbert space $\mathcal{G}$. Therefore, there exists an orthogonal projection operator from $\mathcal{G}$ to $\mathcal{G}_c$ which is linear, continuous, of unit norm, positive, and self-adjoint. $\mathcal{P}_{\mathcal{G}_c}(\mathbf{x})$ is the minimum with respect to the constant vector $\mathbf{a}$ of the integral $\int_\Omega \|\mathbf{x}(\mathbf{r})-\mathbf{a}\|^2\,d\mathbf{r}$. Taking the derivative with respect to $\mathbf{a}$, we obtain the necessary condition

$$\int_\Omega (\mathbf{x}(\mathbf{r})-\mathbf{a})\,d\mathbf{r} = 0$$

and hence $\mathbf{a}_{min} = \overline{\mathbf{x}}$. Conversely, $\mathbf{x}-\mathbf{a}_{min}$ is orthogonal to $\mathcal{G}_c$ since $\int_\Omega(\mathbf{x}(\mathbf{r}) - \mathbf{a}_{min})\mathbf{b}\,d\mathbf{r} = 0$ for all $\mathbf{b}\in\mathcal{G}_c$.

Let $\mathbf{y}\in\mathcal{G}$, $\int_\Omega \mathbf{x}\,\mathbf{y}(\mathbf{r})\,d\mathbf{r} = \mathbf{x}\int_\Omega \mathbf{y}(\mathbf{r})\,d\mathbf{r} = 0$ for all $\mathbf{x}\in\mathcal{G}_c$ if and only if $\mathbf{y}\in\mathcal{G}_c^\perp$.

Finally,

$$\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{M}\mathbf{x} = \mathbf{M}\mathbf{x} - \overline{\mathbf{M}\mathbf{x}} = \mathbf{M}\mathbf{x} - \mathbf{M}\overline{\mathbf{x}} = \mathbf{M}(\mathbf{x}-\overline{\mathbf{x}}) = \mathbf{M}\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{x}. \qquad \square$$

We are now ready to prove the theorem on the absolute stability of the homogeneous solutions to (2.6).

THEOREM 5.2. *If $\mathbf{W}$ satisfies* (3.5)*, a sufficient condition for the absolute stability of a homogeneous solution to* (2.6) *is that the norm $\|g^*\|_{\mathcal{G}_c^\perp}$ of the restriction to $\mathcal{G}_c^\perp$ of the compact operator g^* be less than $(\tau_{\max}DS_m)^{-1}$ for all $t\in$* J.

Proof. This proof is inspired by [31]. Note that $\mathcal{G}_c^\perp$ is invariant by g^* and hence by g_m^*. Indeed, from Lemma 4.4 and (3.5) we have

$$\overline{g^*(\mathbf{x})} = \overline{\mathbf{W}^T}(t)\overline{\mathbf{x}} = 0 \quad \forall\mathbf{x}\in\mathcal{G}_c^\perp.$$

Let $\mathbf{V}_p$ be the unique solution of (2.6) with homogeneous input $\mathbf{I}_{\mathrm{ext}}(t)$ and initial conditions $\mathbf{V}_p(0) = \mathbf{V}_{p0}\in\mathcal{G}$, and consider the initial value problem

$$\left\{\begin{array}{rcl} \mathbf{X}'(t) & = & \mathcal{P}_{\mathcal{G}_c^\perp}\left(f_v(t, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X} + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p)\right), \\ \mathbf{X}(0) & = & \mathbf{X}_0. \end{array}\right. \tag{5.2}$$

$\mathbf{X} = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{V}_p$ is a solution with initial condition $\mathbf{X}_0 = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{V}_{p0}$, since $\mathcal{P}_{\mathcal{G}_c^\perp}^2 = \mathcal{P}_{\mathcal{G}_c^\perp}$, and $\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{V}_p + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p = \mathbf{V}_p$. But $\mathbf{X} = 0$ is also a solution with initial condition $\mathbf{X}_0 = 0$. Indeed $\mathcal{G}_c$ is flow-invariant because of (3.5), that is, $f_v(t,\mathcal{G}_c)\subset\mathcal{G}_c$, and hence $\mathcal{P}_{\mathcal{G}_c^\perp}(f_v(t,\mathcal{G}_c)) = 0$. We therefore look for a sufficient condition for the system (5.2) to be absolutely stable at $\mathbf{X} = 0$.

[2] To be accurate, this is the projection on the closure of $\mathcal{G}_c^\perp$ in the closure of $\mathcal{G}$, which is the Hilbert space $\mathbf{L}_n^2(\Omega)$.

We consider again the functional $V(\mathbf{X}) = \frac{1}{2}\left\langle \mathbf{X},\, \mathbf{L}^{-1}\mathbf{X}\right\rangle$ with time derivative $\frac{dV(\mathbf{X})}{dt} = \left\langle\, \mathbf{X},\, \mathbf{L}^{-1}\mathbf{X}_t \,\right\rangle$. We substitute $\mathbf{X}_t$ with its value from (5.2) which can be rewritten as

$$\mathbf{X}_t = \mathcal{P}_{\mathcal{G}_c^\perp}\left(-\mathbf{L}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X} + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p) + \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}',t)\,\mathbf{S}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t) + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t))\,d\mathbf{r}'\right).$$

Because of Lemma 5.1 this yields

$$\mathbf{X}_t = -\mathbf{L}\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X} + \mathcal{P}_{\mathcal{G}_c^\perp}\left(\int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)\,\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t) + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t))\,d\mathbf{r}'\right).$$

Using a zeroth-order Taylor expansion, as in the proof of Lemma 3.1, we write

$$\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X} + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p) = \underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p) + \left(\int_0^1 D\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p + \zeta\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X})\,d\zeta\right)\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X},$$

and since $\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p) \in \mathcal{G}_c$, and because of (3.5),

$$\begin{aligned}&\mathcal{P}_{\mathcal{G}_c^\perp}\left(\int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)\,\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t) + \mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t))\,d\mathbf{r}'\right)\\ &= \mathcal{P}_{\mathcal{G}_c^\perp}\left(\int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)\left(\int_0^1 D\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t)+\zeta\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t))\,d\zeta\right)\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t)\,d\mathbf{r}'\right).\end{aligned}$$

We use (5.1) and the fact that $\mathcal{P}_{\mathcal{G}_c^\perp}$ is self-adjoint and idempotent to write

$$\begin{aligned}\frac{dV(\mathbf{X})}{dt} = -\langle\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X},\, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}\rangle + \Bigg\langle \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X},\, \mathbf{L}^{-1}\int_\Omega \mathbf{W}_{cm}(\mathbf{r},\mathbf{r}',t)&\\ \cdot\left(\int_0^1 D\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t) + \zeta\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t))\,d\zeta\right)\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t)\,d\mathbf{r}'\Bigg\rangle&.\end{aligned}$$

Let us denote by $\mathcal{D}_v(\mathbf{r}',t)$ the diagonal matrix $\int_0^1 D\underline{\mathbf{S}}(\mathcal{P}_{\mathcal{G}_c}\mathbf{V}_p(\mathbf{r}',t)+\zeta\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}(\mathbf{r}',t))\,d\zeta$. Its diagonal elements are continuous functions with values between 0 and 1. Letting $\mathbf{Y} = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{X}$, we rewrite the previous equation in operator form, introducing the operator g_m (Definition 4.1), as

$$\frac{dV(\mathbf{X})}{dt} = -\langle\mathbf{Y},\, \mathbf{Y}\rangle + \left\langle\mathbf{Y},\, \mathbf{L}^{-1}g_m(\mathcal{D}_v\,\mathbf{Y})\right\rangle.$$

By definition of the adjoint

$$\left\langle\mathbf{Y},\, \mathbf{L}^{-1}g_m(\mathcal{D}_v\,\mathbf{Y})\right\rangle = \left\langle g_m^*\left(\mathbf{L}^{-1}\mathbf{Y}\right),\, \mathcal{D}_v\,\mathbf{Y}\right\rangle.$$

Using the Cauchy–Schwarz inequality and Lemma 4.5

$$\left|\left\langle g_m^*\left(\mathbf{L}^{-1}\mathbf{Y}\right),\, \mathcal{D}_v\,\mathbf{Y}\right\rangle\right| \le \left\|g_m^*\left(\mathbf{L}^{-1}\mathbf{Y}\right)\right\|_{n,2}\|\mathcal{D}_v\,\mathbf{Y}\|_{n,2} \le \left\|g_m^*\left(\mathbf{L}^{-1}\mathbf{Y}\right)\right\|_{n,2}\|\mathbf{Y}\|_{n,2},$$

and since

$$\left\|g_m^*\left(\mathbf{L}^{-1}\mathbf{Y}\right)\right\|_{n,2} \le \|g_m^*\|_{\mathcal{G}_c^\perp}\left\|\mathbf{L}^{-1}\mathbf{Y}\right\|_{n,2} \le \tau_{\max}DS_m\|g^*\|_{\mathcal{G}_c^\perp}\|\mathbf{Y}\|_{n,2},$$

the conclusion follows. □

Note that $\|g^*\|_{\mathcal{G}_c^\perp} = \|g^*\|_{\mathbf{L}_0^2}$ by density of $\mathcal{G}_c^\perp$ in $\mathbf{L}_0^2$, where $\mathbf{L}_0^2$ is the subspace of $\mathbf{L}^2$ of zero mean functions. We show in Appendix A how to compute this norm.

We prove a similar theorem in the case of (2.7).

THEOREM 5.3. *If* $\mathbf{W}$ *satisfies* (3.5), *a sufficient condition for the stability of a homogeneous solution to* (2.7) *is that the norm* $\|g\|_{\mathcal{G}_c^\perp}$ *of the restriction to* $\mathcal{G}_c^\perp$ *of the compact operator* g *be less than* $(\tau_{\max} DS_m)^{-1}$ *for all* $t \in \mathrm{J}$.

Proof. The proof is similar to that of Theorem 5.2. We consider $\mathbf{A}_p$ the unique solution to (2.7) with homogeneous input $\mathbf{I}_{\text{ext}}(t)$ and initial conditions $\mathbf{A}_p(0) = \mathbf{A}_{p0}$, and we consider the initial value problem

$$\left\{ \begin{array}{lcl} \mathbf{A}'(t) & = & \mathcal{P}_{\mathcal{G}_c^\perp}\left(f_a(t, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A} + \mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p)\right), \\ \mathbf{A}(0) & = & \mathbf{A}_0. \end{array} \right. \tag{5.3}$$

$\mathbf{A} = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}_p$ is a solution with initial conditions $\mathbf{A}_0 = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}_{p0}$ since $\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}_p + \mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p = \mathbf{A}_p$. But $\mathbf{A} = 0$ is also a solution with initial conditions $\mathbf{A}_0 = 0$. Indeed, $\mathcal{G}_c$ is flow-invariant because of (3.5), that is, $f_a(t, \mathcal{G}_c) \subset \mathcal{G}_c$, and hence $\mathcal{P}_{\mathcal{G}_c^\perp}(f_a(t, \mathcal{G}_c)) = 0$. We therefore look for a sufficient condition for the system (5.3) to be absolutely stable at $\mathbf{A} = 0$.

Consider again the functional $V(\mathbf{A}) = \frac{1}{2}\left\langle \mathbf{A}, \mathbf{L}^{-1}\mathbf{A} \right\rangle$ with time derivative $\frac{dV(\mathbf{A})}{dt} = \left\langle \mathbf{A}, \mathbf{L}^{-1}\mathbf{A}_t \right\rangle$. We substitute $\mathbf{A}_t$ with its value from (5.3) which, using (3.5), can be rewritten as

$$\mathbf{A}_t = \mathcal{P}_{\mathcal{G}_c^\perp}\Bigg(-\mathbf{L}(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A} + \mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p) + \mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}(\mathbf{r}', t)\, d\mathbf{r}' + \overline{\mathbf{W}}(t)\mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p + \mathbf{I}_{\text{ext}}(t) \right) \Bigg).$$

We perform a first-order Taylor expansion with integral remainder of the $\mathbf{S}$ term and introduce the operator h_m (Definition 4.1):

$$\begin{aligned} &\mathbf{S}\left(\int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}(\mathbf{r}', t)\, d\mathbf{r}' + \overline{\mathbf{W}}(t)\mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p + \mathbf{I}_{\text{ext}}(t) \right) = \\ &\qquad \mathbf{S}\left(\overline{\mathbf{W}}(t)\mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p + \mathbf{I}_{\text{ext}}(t) \right) \\ &+ \left(\int_0^1 D\underline{\mathbf{S}}\left(\overline{\mathbf{W}}(t)\mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p + \mathbf{I}_{\text{ext}}(t) + \zeta \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}(\mathbf{r}', t)\, d\mathbf{r}' \right) d\zeta \right) h_m(\mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A})(\mathbf{r}, t). \end{aligned}$$

Let us define

$$\mathcal{D}_a(\mathbf{r}, t) = \int_0^1 D\underline{\mathbf{S}}\left(\overline{\mathbf{W}}(t)\mathcal{P}_{\mathcal{G}_c}\mathbf{A}_p + \mathbf{I}_{\text{ext}}(t) + \zeta \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\, \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}(\mathbf{r}', t)\, d\mathbf{r}' \right) d\zeta,$$

a diagonal matrix whose diagonal elements are continuous functions with values between 0 and 1. Letting $\mathbf{Y} = \mathcal{P}_{\mathcal{G}_c^\perp}\mathbf{A}$, we write

$$\frac{dV(\mathbf{A})}{dt} = -\langle \mathbf{Y}, \mathbf{Y} \rangle + \left\langle \mathbf{Y}, \mathbf{L}^{-1}\mathcal{D}_a\, h_m(\mathbf{Y}) \right\rangle,$$

and the conclusion follows from the Cauchy–Schwarz inequality and Lemmas 4.5 and 4.6:

$$\begin{aligned}\left|\left\langle \mathbf{Y},\, \mathbf{L}^{-1}\mathcal{D}_a\, h_m(\mathbf{Y})\right\rangle\right| &\leq\leq \|\mathbf{Y}\|_{n,2}\, \left\|\mathbf{L}^{-1}\mathcal{D}_a\, h_m(\mathbf{Y})\right\|_{n,2} \\ &\leq \tau_{\max}\, \|\mathbf{Y}\|_{n,2}\, \|h_m(\mathbf{Y})\|_{n,2} \leq \tau_{\max} DS_m\, \|g\|_{\mathcal{G}_c^{\perp}}\, \|\mathbf{Y}\|_{n,2}^{2}. \qquad \square\end{aligned}$$

5.2. The convolution case. As $\mathbf{W}$ is translation-invariant, $\int_\Omega \mathbf{W}(\mathbf{r}-\mathbf{r}',t)\,d\mathbf{r}'$ is in general a function of $\mathbf{r}$, unless Ω has no border. In our framework, this case occurs only when Ω is an interval with periodic conditions and we have the following theorem.

THEOREM 5.4. *A sufficient condition for the stability of a homogeneous solution to* (2.6) *(resp.,* (2.7)*) is that the eigenvalues of the Hermitian matrices*

$$\widetilde{\mathbf{W}}^*(\mathbf{m},t)\widetilde{\mathbf{W}}(\mathbf{m},t)$$

are strictly less than $(\tau_{\max}\, DS_m)^{-2}$ *for all* $\mathbf{m} \neq \mathbf{0} \in \mathbb{Z}^q$ *and all* $t \in \mathrm{J}$. $\widetilde{\mathbf{W}}(\mathbf{m},t)$ *is the* $\mathbf{m}$*th element of the Fourier series of the q-periodic matrix function* $\mathbf{W}(\mathbf{r},t)$ *with respect to the space variable* $\mathbf{r}$.

The only difference from Theorem 4.12 is that there are no constraints on the Fourier coefficient $\mathbf{m} = 0$. This is due to the fact that we "look" only at the subspace of $\mathcal{G}$ of functions with zero spatial average.

5.3. Complete synchronization. The property of absolute stability of the solution that is characterized in Theorems 5.2, 5.3, and 5.4 can be seen as the ability for the neural masses in the continuum to synchronize. By synchronization we mean that the state vectors at all points in the continuum converge to a unique state vector that is a function only of the common input $\mathbf{I}_{\text{ext}}$ and not of the initial states of the neural masses. The state vector is the homogeneous solution of (2.6) and (2.7). This effect is called complete synchronization [32].

6. Extending the theory. We have developed our analysis of (2.6) and (2.7) in the Banach space $\mathcal{F}$ of continuous functions of the spatial coordinate $\mathbf{r}$ even though we have used a structure of pre-Hilbert space $\mathcal{G}$ on top of it. But there remains the fact that the solutions that we have been discussing are smooth, i.e., continuous with respect to the space variable. It may be interesting to also consider nonsmooth solutions, e.g., piecewise continuous solutions that can be discontinuous along curves of Ω. A natural setting, given the fact that we are interested in having a structure of Hilbert space, is $\mathbf{L}^2_n(\Omega)$, the space of square-integrable functions from Ω to $\mathbb{R}^n$; see Appendix A. It is a Hilbert space and $\mathcal{G}$ is a dense subspace, $\overline{\mathcal{G}} = \mathbf{L}^2_n(\Omega)$, where $\overline{A}$ indicates the topological closure of the set A.

6.1. Existence, uniqueness, and stability of a solution. The theory developed in the previous sections can be readily extended to $\mathbf{L}^2_n(\Omega)$: the analysis of the stability of the general and homogeneous solutions has been done using the pre-Hilbert space structure of $\mathcal{G}$, and all the operators that have been shown to be compact in $\mathcal{G}$ are also compact in its closure $\mathbf{L}^2_n(\Omega)$ [9]. The only point that has to be reworked is the problem of existence and uniqueness of a solution addressed in Propositions 3.2 and 3.3. This allows us to *relax* the rather stringent spatial smoothness hypotheses imposed on the connectivity function $\mathbf{W}$ and the external current $\mathbf{I}_{\text{ext}}$, thereby bringing in more flexibility to the model. We have the following proposition.

PROPOSITION 6.1. *If the following two hypotheses are satisfied:*

1. *the mapping* $\mathbf{W}$ *is in* $C(\mathrm{J}; \mathbf{L}^2_{\mathrm{n}\times\mathrm{n}}(\Omega \times \Omega))$*,*
2. *the external current* $\mathbf{I}_{\mathrm{ext}}$ *is in* $C(\mathrm{J}; \mathbf{L}^2_{\mathrm{n}}(\Omega))$*,*

then the mappings f_v *and* f_a *are from* $\mathrm{J} \times \mathbf{L}^2_{\mathrm{n}}(\Omega)$ *to* $\mathbf{L}^2_n(\Omega)$*, continuous, and Lipschitz continuous with respect to their second argument, uniformly with respect to the first.*

Proof. Because of the first hypothesis, the fact that $\mathbf{S}(\mathbf{x})$ is in $\mathbf{L}^2_n(\Omega)$ for all $\mathbf{x} \in \mathbf{L}^2_n(\Omega)$, and because of Lemma A.2, f_v is well defined. Let us prove that it is continuous. As in the proof of Proposition 3.2, we write

$$f_v(t,\mathbf{x}) - f_v(s,\mathbf{y}) = -\mathbf{L}(\mathbf{x}-\mathbf{y}) + \int_\Omega (\mathbf{W}(\cdot,\mathbf{r}',t) - \mathbf{W}(\cdot,\mathbf{r}',s))\mathbf{S}(\mathbf{x}(\mathbf{r}'))\,d\mathbf{r}'$$
$$+ \int_\Omega \mathbf{W}(\cdot,\mathbf{r}',s)(\mathbf{S}(\mathbf{x}(\mathbf{r}')) - \mathbf{S}(\mathbf{y}(\mathbf{r}')))\,d\mathbf{r}' + \mathbf{I}_{\mathrm{ext}}(\cdot,t) - \mathbf{I}_{\mathrm{ext}}(\cdot,s),$$

from which we obtain, using Lemma A.2,

$$\|f_v(t,\mathbf{x}) - f_v(s,\mathbf{y})\|_{n,2} \le \|\mathbf{L}\|_F\|\mathbf{x}-\mathbf{y}\|_{n,2} + \sqrt{n|\Omega|}S_m\|\mathbf{W}(\cdot,\cdot,t) - \mathbf{W}(\cdot,\cdot,s)\|_F$$
$$+ DS_m\|\mathbf{W}(\cdot,\cdot,s)\|_F\|\mathbf{x}-\mathbf{y}\|_{n,2} + \|\mathbf{I}_{\mathrm{ext}}(\cdot,t) - \mathbf{I}_{\mathrm{ext}}(\cdot,s)\|_{n,2},$$

and the continuity follows from the hypotheses. $\|\ \|_F$ is the Frobenius norm; see Appendix A. Note that since $\mathbf{W}$ is continuous on the compact interval J, it is bounded and $\|\mathbf{W}(\cdot,\cdot,t)\|_F \le w$ for all $t \in \mathrm{J}$ for some positive constant w. The Lipschitz continuity with respect to the second argument uniformly with respect to the first one follows from the previous inequality by choosing $s = t$.

The proof for f_a is similar. □

From this proposition we deduce the existence and uniqueness of a solution over a subinterval of $\mathbb{R}$.

PROPOSITION 6.2. *Subject to the hypotheses of Proposition* 6.1 *for any element* $\mathbf{V}_0$ *of* $\mathbf{L}^2_n(\Omega)$ *there is a unique solution* $\mathbf{V}$*, defined on a subinterval of* J *containing* 0 *and continuously differentiable, of the abstract initial value problem* (3.1) *for* $f = f_v$ *and* $f = f_a$ *such that* $\mathbf{V}(0) = \mathbf{V}_0$*.*

Proof. All conditions of the Picard–Lindelöf theorem on differential equations in Banach spaces (here a Hilbert space) [9, 2] are satisfied; hence the proposition is proved. □

We can also prove that this solution exists for all times, as in Proposition 3.4.

PROPOSITION 6.3. *If the following two hypotheses are satisfied:*

1. *the connectivity function* $\mathbf{W}$ *is in* $C(\mathbb{R}; \mathbf{L}^2_{\mathrm{n}\times\mathrm{n}}(\Omega \times \Omega))$*,*
2. *the external current* $\mathbf{I}_{\mathrm{ext}}$ *is in* $C(\mathbb{R}; \mathbf{L}^2_{\mathrm{n}}(\Omega))$*,*

then for any function $\mathbf{V}_0$ *in* $\mathbf{L}^2_n(\Omega)$ *there is a unique solution* $\mathbf{V}$*, defined on* $\mathbb{R}$ *and continuously differentiable, of the abstract initial value problem* (3.1) *for* $f = f_v$ *and* $f = f_a$*.*

Proof. The proof is similar to that of Proposition 3.4. □

The absolute stability of the solution can be studied exactly as in Theorems 4.7 and 4.8. Since $\mathcal{G}$ is dense in $\mathbf{L}^2_n(\Omega)$, we have $\|g\|_{\mathcal{G}} = \|g\|_{\mathbf{L}^2_n(\Omega)}$ and similar relations for all the other operators. We have the following theorem.

THEOREM 6.4. *If the compact operator* g *satisfies the condition of Theorem* 4.7*, the solution of the abstract initial value problem* (3.1) *for* $f = f_v$ *and* $f = f_a$ *is absolutely stable.*

6.2. Locally homogeneous solutions. An application of the previous extension is the following. Consider a partition of Ω into P subregions Ω_i, $i = 1, \ldots, P$.

We assume that the Ω_is are closed, hence compact, subsets of Ω intersecting along finitely many piecewise regular curves. These curves form a set of 0 Lebesgue measure of Ω. We consider locally homogeneous input current functions

$$\mathbf{I}_{\text{ext}}(\mathbf{r},t) = \sum_{k=1}^{P} \mathbf{1}_{\Omega_k}(\mathbf{r})\mathbf{I}_{\text{ext}}^k(t), \tag{6.1}$$

where the P functions $\mathbf{I}_{\text{ext}}^k(t)$ are continuous on some closed interval J containing 0. On the border between two adjacent regions, the value of $\mathbf{I}_{\text{ext}}(\mathbf{r},t)$ is undefined. Since this set of borders is of 0 measure, the functions defined by (6.1) are in $\mathbf{L}_n^2(\Omega)$ at each time instant.

6.2.1. Existence and uniqueness. We are interested in the existence of solutions to the abstract initial value problem (3.1) that are homogeneous in each subregion Ω_i, $i = 1, \ldots, P$. We call them locally homogeneous solutions.

We assume that the connectivity matrix $\mathbf{W}$ satisfies the following conditions:

$$\int_{\Omega_k} \mathbf{W}(\mathbf{r},\mathbf{r}',t)\, d\mathbf{r}' = \sum_{i=1}^{P} \mathbf{1}_{\Omega_i}(\mathbf{r})\mathbf{W}_{ik}(t), \quad k = 1, \ldots, P. \tag{6.2}$$

These conditions are analogous to (3.5). A locally homogeneous solution of (2.6) or (2.7) can be written as

$$\mathbf{V}(\mathbf{r},t) = \sum_{i=1}^{P} \mathbf{1}_{\Omega_i}(\mathbf{r})\mathbf{V}_i(t),$$

where the functions $\mathbf{V}_i$ satisfy the system of ordinary differential equations

$$\mathbf{V}_i'(t) = -\mathbf{L}\mathbf{V}_i(t) + \sum_{k=1}^{P} \mathbf{W}_{ik}(t)\mathbf{S}(\mathbf{V}_k(t)) + \mathbf{I}_{\text{ext}}^i(t) \tag{6.3}$$

for the voltage-based model and

$$\mathbf{V}_i'(t) = -\mathbf{L}\mathbf{V}_i(t) + \mathbf{S}\left(\sum_{k=1}^{P} \mathbf{W}_{ik}(t)\mathbf{V}_k(t) + \mathbf{I}_{\text{ext}}^i(t)\right) \tag{6.4}$$

for the activity-based model. The conditions for the existence and uniqueness of a locally homogeneous solution are given in the following theorem, analogous to Theorem 3.5.

THEOREM 6.5. *If the external currents $\mathbf{I}_{\text{ext}}^k(t)$, $k = 1, \ldots, P$, and the connectivity matrices $\mathbf{W}_{ik}(t)$, $i, k = 1, \ldots, P$, are continuous on some closed interval* J *containing* 0*, then for all sets of P vectors $\mathbf{U}_0^k$, $k = 1, \ldots, P$, of $\mathbb{R}^n$, there exists a unique solution $(\mathbf{U}_1(t), \ldots, \mathbf{U}_P(t))$ of* (6.3) *or* (6.4) *defined on a subinterval* J_0 *of* J *containing* 0 *such that $\mathbf{U}_k(0) = \mathbf{U}_0^k$, $k = 1, \ldots, P$.*

Proof. The system (6.3) can be written in the form

$$\mathcal{V}'(t) = -\mathcal{L}\mathcal{V}(t) + \mathcal{W}(t)\mathcal{S}(\mathcal{V}(t)) + \mathcal{I}_{\text{ext}}(t), \tag{6.5}$$

where $\mathcal{V}$ is the nP-dimensional vector

$$\begin{pmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_P \end{pmatrix}, \mathcal{I}_{\text{ext}} = \begin{pmatrix} \mathbf{I}_{\text{ext}}^1 \\ \vdots \\ \mathbf{I}_{\text{ext}}^P \end{pmatrix}, \mathcal{S}(\mathcal{X}) = \begin{pmatrix} \mathbf{S}(\mathbf{X}_1) \\ \vdots \\ \mathbf{S}(\mathbf{X}_P) \end{pmatrix},$$

$\mathcal{W}$ is the block matrix $(\mathbf{W}_{ik})_{i,k}$, and $\mathcal{L}$ is the block diagonal matrix whose diagonal elements are all equal to $\mathbf{L}$. Then we are dealing with a classical initial value problem of dimension nP, and the proof of existence and uniqueness is similar to that of Theorem 3.5. A similar proof can be written in the case of system (6.4). $\square$

Again, if $\mathcal{I}_{\text{ext}}$ and $\mathcal{W}$ are continuous on $\mathbb{R}$, the existence and uniqueness result extends to the whole time line $\mathbb{R}$.

6.2.2. Absolute stability. Having proved the existence and uniqueness of a locally homogeneous solution, we consider the problem of characterizing its absolute stability. The method is the same as in section 5. We consider the subset, noted $\mathcal{G}_c^P$, of the functions that are constant in the interior $\overset{\circ}{\Omega}_i$ of each region $\Omega_i, i = 1, \dots, P$ (the interior $\overset{\circ}{A}$ of a subset A is defined as the biggest open subset included in A). We have the following lemma that echoes Lemma 5.1.

Lemma 6.6. *$\mathcal{G}_c^P$ is a complete linear subspace of $\mathbf{L}_n^2(\Omega)$. The orthogonal projection operator $\mathcal{P}_{\mathcal{G}_c^P}$ from $\mathbf{L}_n^2(\Omega)$ to $\mathcal{G}_c^P$ is defined by*

$$\mathcal{P}_{\mathcal{G}_c^P}(\mathbf{x})(\mathbf{r}) = \overline{\mathbf{x}}^P = \sum_{k=1}^{P} \mathbf{1}_{\Omega_k}(\mathbf{r}) \frac{1}{|\Omega_k|} \int_{\Omega_k} \mathbf{x}(\mathbf{r}')\, d\mathbf{r}'.$$

The orthogonal complement $\mathcal{G}_c^{P\,\perp}$ of $\mathcal{G}_c^P$ is the subset of functions of $\mathbf{L}_n^2(\Omega)$ that have a zero average in each Ω_i, $i = 1, \dots, P$. The orthogonal projection operator $\mathcal{P}_{\mathcal{G}_c^{P\,\perp}}$ is equal to $\mathrm{Id} - \mathcal{P}_{\mathcal{G}_c^{\mathrm{P}}}$. *We also have*

$$\mathcal{P}_{\mathcal{G}_c^{P\,\perp}} \mathbf{M}\mathbf{x} = \mathbf{M}\mathcal{P}_{\mathcal{G}_c^{P\,\perp}}\mathbf{x} \quad \forall\, \mathbf{x} \in \mathbf{L}_n^2(\Omega),\, \mathbf{M} \in \mathcal{M}_{n\times n}. \tag{6.6}$$

Proof. The proof of this lemma is similar to that of Lemma 5.1. $\square$

We have the following theorem, corresponding to Theorems 5.2 and 5.3.

Theorem 6.7. *If $\mathbf{W}$ satisfies* (6.2)*, a sufficient condition for the absolute stability of a locally homogeneous solution to* (2.6) *(resp.,* (2.7)*) is that the norm $\|g^*\|_{\mathcal{G}_c^{P\,\perp}}$ (resp., $\|g\|_{\mathcal{G}_c^{P\,\perp}}$) of the restriction to $\mathcal{G}_c^{P\,\perp}$ of the compact operator g^* (resp., g) be less than $(\tau_{\max} DS_m)^{-1}$ for all $t \in \mathrm{J}$.*

Proof. The proof strictly follows the lines of the ones of Theorems 5.2 and 5.3. $\square$

Note that the condition on the operator norm in Theorems 4.7 and 4.8 is stronger than the one of Theorems 5.2 and 5.3 which is in turn stronger than the one of Theorem 6.7; therefore, we have the following proposition.

Proposition 6.8. *If the operator g satisfies the condition of Theorem* 4.7 *or if $g*$ (resp., g) satisfies the condition of Theorem* 5.2 *(resp., of Theorem* 5.3*), then for every partition of Ω, corresponding locally homogeneous current, and $\mathbf{W}$ satisfying* (6.2)*, the locally homogeneous solution of* (2.6) *(resp.,*(2.7)*) is absolutely stable.*

Proof. Since all spaces are contained in $\mathbf{L}_n^2(\Omega)$, the first part of the proposition is proved. Next, it is clear that $\mathcal{G}_c \subset \mathcal{G}_c^P$; therefore, $\mathcal{G}_c^{P\,\perp} \subset \mathcal{G}_c^{\perp}$ and $\|g^*\|_{\mathcal{G}_c^{P\,\perp}} \le \|g^*\|_{\mathcal{G}_c^{\perp}}$ (resp., $\|g\|_{\mathcal{G}_c^{P\,\perp}} \le \|g\|_{\mathcal{G}_c^{\perp}}$). $\square$

Condition (6.2) depends on the partition of Ω. It is therefore unrealistic since one expects this partition to change over time with the external currents. In this context it is interesting to define the notion of a *pseudo–locally homogeneous* solution.

Definition 6.9. *A pseudo–locally homogeneous solution of* (2.6) *(resp.,* (2.7)*) corresponds to a locally homogeneous input current (verifying* (6.1)*) when the connectivity function satisfies the condition of Proposition* 6.3 *(existence and uniqueness of a solution) but not necessarily conditions* (6.2).

How much a pseudo–locally homogeneous solution differs from a locally homogeneous solution obviously depends upon how poorly the connectivity function satisfies the conditions (6.2). But since pseudo–locally homogeneous solutions are solutions, they enjoy the following property.

PROPOSITION 6.10. *If the operator g satisfies the condition of Theorem* 4.7*, the unique pseudo–locally homogeneous solution of* (2.6) *(resp., of* (2.7)*) corresponding to a locally homogeneous input current is absolutely stable.*

A numerical example of pseudo–locally homogeneous solution is given in section 7 (Figures 17 and 18).

6.2.3. Complete local synchronization. The property of absolute stability of the solution that is characterized in Theorem 6.7 can be seen as the ability for the neural masses in the continuum to synchronize locally within each region Ω_i, $i = 1, \dots, P$. By local synchronization we mean that the state vectors at all points of each region Ω_i converge to a unique state vector that is a function only of the common input $\mathbf{I}^i_{\text{ext}}$ within Ω_i and not of the initial states of the neural masses. The state vector is the locally homogeneous solution of (2.6) and (2.7). This effect is called complete local synchronization.

6.3. Higher order postsynaptic potentials. We now show how the theory developed so far can be extended to accomodate more complicated time variations of the postsynaptic potentials than the decaying exponential that we adopted so far with the advantage that we had only to deal with a first-order differential equation. We show how to proceed only in the case of a second-order differential equation; going to a higher order does not bring in new difficulties. We also treat only the case of the voltage-based model, the case of the activity-based model being similar.

We therefore assume that, with the notation of section 2.1.1, we have $PSP_i(t) = te^{-t/\tau_i}Y(t)$ or, equivalently, that

$$\frac{d^2 PSP_i(t)}{dt^2} + \frac{2}{\tau_i}\frac{dPSP_i(t)}{dt} + \frac{1}{\tau_i^2} = \delta(t).$$

The analogue of (2.4) is

$$\mathbf{V}'' = -2\mathbf{L}\mathbf{V}' - \mathbf{L}^2\mathbf{V} + \mathbf{W}\,\mathbf{S}(\mathbf{V}) + \mathbf{I}_{\text{ext}}. \tag{6.7}$$

We rewrite this as a first order system of differential equations by introducing the vector $\boldsymbol{\mathcal{V}} = \left[\begin{smallmatrix}\mathbf{V}\\ \mathbf{V}'\end{smallmatrix}\right]$:

$$\boldsymbol{\mathcal{V}}' = -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{V}} + \begin{bmatrix}\mathbf{0}\\ \mathbf{W}\,\mathbf{S}(\mathbf{V})\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{\text{ext}}\end{bmatrix}, \quad \boldsymbol{\mathcal{L}} = \begin{bmatrix}\mathbf{0} & -\text{Id}\\ \mathbf{L}^2 & 2\mathbf{L}\end{bmatrix}.$$

The dynamic system $\boldsymbol{\mathcal{V}}' = -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{V}}$ is globally asymptotically stable since all the eigenvalues of the $2n \times 2n$ matrix $\boldsymbol{\mathcal{L}}$ have a strictly positive real part, as can be easily verified.[3] This has the following consequence [36, 27] that is used below and that we cite without proof.

THEOREM 6.11 (Lyapunov). *The symmetric positive definite matrix*

$$\boldsymbol{\mathcal{M}} = \int_0^\infty e^{-\boldsymbol{\mathcal{L}}^T t}\, e^{-\boldsymbol{\mathcal{L}} t}\, dt$$

[3] In fact, the eigenvalues of $\boldsymbol{\mathcal{L}}$ are those of $\mathbf{L}$, $1/\tau_i$s, with multiplicity 2.

satisfies

$$\boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{L}}^T\boldsymbol{\mathcal{M}} = \mathrm{Id}_{2n}, \tag{6.8}$$

where Id_{2n} *is the* $2n \times 2n$ *identity matrix.*

The analogue of (2.6) is readily found to be

$$\boldsymbol{\mathcal{V}}_t(\mathbf{r}, t) = -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{V}}(\mathbf{r}, t) + \left[\begin{array}{c} \mathbf{0} \\ \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}', t)\mathbf{S}(\mathbf{V}(\mathbf{r}', t))\, d\mathbf{r}' \end{array} \right] . + \left[\begin{array}{c} \mathbf{0} \\ \mathbf{I}_{\mathrm{ext}} \end{array} \right]. \tag{6.9}$$

The state is now $2n$-dimensional, the corresponding functional space is $\mathbf{L}^2_{2n}(\Omega)$, and the operator g is defined on the subspace $\mathbf{L}^2_n(\Omega)$ of $\mathbf{L}^2_{2n}(\Omega)$. It keeps all its previous properties. All proofs of the existence and uniqueness of a solution to (2.6) extend mutatis mutandis to this new setting.

Let us now examine the problem of the absolute stability of the solution, the analogue of Theorem 4.7.

Theorem 6.12. *A sufficient condition for the solution of* (2.6) *to be absolutely stable is*

$$2\, \lambda_{\max}\, DS_m\, \|g\|_{\mathbf{L}^2_n(\Omega)} < 1,$$

where $\lambda_{\max}$ *is the largest eigenvalue of the* $2n \times 2n$ *matrix* $\boldsymbol{\mathcal{M}}$ *defined in Theorem* 6.11.

Proof. We consider the equation

$$\boldsymbol{\mathcal{V}}_t(\mathbf{r}, t) = -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{V}}(\mathbf{r}, t) + \left[\begin{array}{c} \mathbf{0} \\ \int_\Omega \mathbf{W}_{cm}(\mathbf{r}, \mathbf{r}', t)\underline{\mathbf{S}}(\mathbf{V}(\mathbf{r}', t))\, d\mathbf{r}' \end{array} \right] + \left[\begin{array}{c} \mathbf{0} \\ \mathbf{I}_{\mathrm{ext}} \end{array} \right],$$

where $\mathbf{V}$ is the vector composed of the first n components of vector $\boldsymbol{\mathcal{V}}$ (the same convention will be used in the following for subvectors of $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{X}}$). Let $\boldsymbol{\mathcal{U}}$ be its unique solution with initial condition $\boldsymbol{\mathcal{U}}(0) = \boldsymbol{\mathcal{U}}_0$, an element of $\mathbf{L}^2_{2n}(\Omega)$. Let also $\boldsymbol{\mathcal{V}}$ be the unique solution of the same equation with different initial conditions $\boldsymbol{\mathcal{V}}(0) = \boldsymbol{\mathcal{V}}_0$, another element of $\mathbf{L}^2_{2n}(\Omega)$. We introduce the new function $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{V}} - \boldsymbol{\mathcal{U}}$ which satisfies

$$\begin{aligned} \boldsymbol{\mathcal{X}}_t(\mathbf{r}, t) &= -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{X}}(\mathbf{r}, t) + \left[\begin{array}{c} \mathbf{0} \\ \int_\Omega \mathbf{W}_{cm}(\mathbf{r}, \mathbf{r}', t)\, \mathbf{H}(\mathbf{X}, \mathbf{U})(\mathbf{r}', t)\, d\mathbf{r}' \end{array} \right] \\ &= -\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{X}}(\mathbf{r}, t) + \left[\begin{array}{c} \mathbf{0} \\ g_m(\mathbf{H}(\mathbf{X}, \mathbf{U}))(\mathbf{r}, t) \end{array} \right], \end{aligned} \tag{6.10}$$

where the vector $\mathbf{H}(\mathbf{X}, \mathbf{U})$ is given by $\mathbf{H}(\mathbf{X}, \mathbf{U})(\mathbf{r}, t) = \underline{\mathbf{S}}(\mathbf{V}(\mathbf{r}, t)) - \underline{\mathbf{S}}(\mathbf{U}(\mathbf{r}, t)) = \underline{\mathbf{S}}(\mathbf{X}(\mathbf{r}, t) + \mathbf{U}(\mathbf{r}, t)) - \underline{\mathbf{S}}(\mathbf{U}(\mathbf{r}, t))$. Consider now the functional

$$V(\boldsymbol{\mathcal{X}}) = \frac{1}{2}\, \langle\, \boldsymbol{\mathcal{X}},\, \boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{X}}\, \rangle,$$

where the symmetric positive definite matrix $\boldsymbol{\mathcal{M}}$ can be seen as defining a metric on the state space. Its time derivative is $\langle\, \boldsymbol{\mathcal{X}},\, \boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{X}}_t\, \rangle$. We replace $\boldsymbol{\mathcal{X}}_t$ by its value from (6.10) in this expression to obtain

$$\frac{dV(\boldsymbol{\mathcal{X}})}{dt} = -\frac{1}{2} \left\langle \boldsymbol{\mathcal{X}},\, (\boldsymbol{\mathcal{L}}^T\boldsymbol{\mathcal{M}} + \boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{L}})\boldsymbol{\mathcal{X}} \right\rangle + \left\langle \boldsymbol{\mathcal{X}},\, \boldsymbol{\mathcal{M}} \left[\begin{array}{c} \mathbf{0} \\ g_m(\mathbf{H}(\mathbf{X}, \mathbf{U})) \end{array} \right] \right\rangle.$$

Using the property (6.8) of $\boldsymbol{\mathcal{M}}$, we obtain

$$\frac{dV(\boldsymbol{\mathcal{X}})}{dt} = -\frac{1}{2}\, \langle\, \boldsymbol{\mathcal{X}},\, \boldsymbol{\mathcal{X}}\, \rangle + \left\langle \boldsymbol{\mathcal{X}},\, \boldsymbol{\mathcal{M}} \left[\begin{array}{c} \mathbf{0} \\ g_m(\mathbf{H}(\mathbf{X}, \mathbf{U})) \end{array} \right] \right\rangle.$$

We consider the second term on the right-hand side of this equation. Since $\boldsymbol{\mathcal{M}}$ is symmetric,

$$\begin{aligned}(6.11)\quad \left|\left\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}\begin{bmatrix}\mathbf{0}\\ g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\end{bmatrix}\right\rangle\right| &= \left|\left\langle \boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{X}}, \begin{bmatrix}\mathbf{0}\\ g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\end{bmatrix}\right\rangle\right|\\ &\leq \|\boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{X}}\|_{2n,2}\,\|g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\|_{n,2} \leq \lambda_{\max}\|\boldsymbol{\mathcal{X}}\|_{2n,2}\,\|g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\|_{n,2}\\ &\leq \lambda_{\max}\,\|\boldsymbol{\mathcal{X}}\|_{2n,2}\,\|g_m\|_{\mathbf{L}^2_n}\|\mathbf{H}(\mathbf{X},\mathbf{U})\|_{n,2}.\end{aligned}$$

The inequality $\|\boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{X}}\|_{2n,2} \leq \lambda_{\max}\|\boldsymbol{\mathcal{X}}\|_{2n,2}$ is obtained using the spectral properties of the symmetric positive definite matrix $\boldsymbol{\mathcal{M}}$ and Lemma 4.5.

Using the idea in the proof of Lemma 3.1, we write $\mathbf{H}(\mathbf{X},\mathbf{U}) = \mathcal{D}_m\mathbf{X}$, where $\mathcal{D}_m$ is a diagonal matrix whose diagonal elements are continuous functions with values between 0 and 1. Hence, because of Lemma 4.5,

$$\|\mathbf{H}(\mathbf{X},\mathbf{U})\|_{n,2} = \|\mathcal{D}_m\mathbf{X}\|_{n,2} \leq \|\mathbf{X}\|_{n,2} \leq \|\boldsymbol{\mathcal{X}}\|_{2n,2}.$$

We use this result and Lemma 4.6 in (6.11) to obtain

$$\left|\left\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{M}}\begin{bmatrix}\mathbf{0}\\ g_m(\mathbf{H}(\mathbf{X},\mathbf{U}))\end{bmatrix}\right\rangle\right| \leq \lambda_{\max}\,DS_m\,\|g\|_{\mathbf{L}^2_n}\,\|\boldsymbol{\mathcal{X}}\|^2_{2n,2},$$

and the conclusion follows. □

All other theorems in sections 4, 5, and 6 and in this section can be similarly extended to this more general setting. More information about $\boldsymbol{\mathcal{M}}$ and $\lambda_{\max}$ can be found in Appendix C.

7. Numerical examples. We consider two ($n = 2$) one-dimensional ($q = 1$) populations of neurons, population 1 being excitatory and population 2 inhibitory. The set Ω is the closed interval $[0, 1]$. We denote x as the spatial variable and f as the spatial frequency variable. We consider Gaussian functions, denoted $G_{ij}(x)$, $i, j = 1, 2$, from which we define the connectivity functions. Hence we have $G_{ij} = \mathcal{G}(0, \sigma_{ij})$. We consider three cases. In the first case, section 7.1, we assume that the connectivity matrix is translation invariant (see sections 4.2 and 5.2). In the second case, section 7.2, we relax this assumption and study the stability of the homogeneous solutions. The third case, finally, section 7.3, covers the case of the locally homogeneous solutions and their stability. In this section we have $S_1(x) = S_2(x) = 1/(1 + e^{-x})$. Therefore,

$$D\mathbf{S}_m = \begin{bmatrix}\frac{1}{4} & 0\\ 0 & \frac{1}{4}\end{bmatrix};$$

hence $DS_m = 1/4$. We also choose $\tau_1 = \tau_2 = 4$; therefore, $\tau_{\max} = 4$, and the product $DS_m\,\tau_{\max}$ is equal to 1.

7.1. The convolution case. We define $W_{ij}(x, x') = \pm\alpha_{ij}\,G_{ij}(x - x')$, where the α_{ij}'s are positive weights and the sign determines whether population j excites (+) or inhibits (−) population i. As explained in section 4.2, $\mathbf{W}(\mathbf{r})$ is defined on the closed interval $\widehat{\Omega} = [-1, 1]$. For simplicity we use the approach described in section 4.2.1 and approximate the Fourier transform of $\mathbf{1}_{\widehat{\Omega}}(x)\mathbf{W}(x)$ by that of $\mathbf{W}(x)$ for which we have an analytical formula. This approximation is good as long as the σ_{ij}'s are small with respect to 1.

The connectivity functions and their Fourier transforms are then given by

$$W_{ij}(x) = \pm \frac{\alpha_{ij}}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{x^2}{2\sigma_{ij}^2}}, \quad \widetilde{W}_{ij}(f) = \pm\alpha_{ij} e^{-2\pi^2 f^2 \sigma_{ij}^2}.$$

The matrices $\mathbf{W}(x)$ and $\widetilde{\mathbf{W}}(f)$ can be written

$$\mathbf{W}(x) = \begin{bmatrix} \frac{\alpha_{11}}{\sqrt{2\pi\sigma_{11}^2}} e^{-\frac{x^2}{2\sigma_{11}^2}} & -\frac{\alpha_{12}}{\sqrt{2\pi\sigma_{12}^2}} e^{-\frac{x^2}{2\sigma_{12}^2}} \\ \frac{\alpha_{21}}{\sqrt{2\pi\sigma_{21}^2}} e^{-\frac{x^2}{2\sigma_{21}^2}} & -\frac{\alpha_{22}}{\sqrt{2\pi\sigma_{22}^2}} e^{-\frac{x^2}{2\sigma_{22}^2}} \end{bmatrix},$$

$$\widetilde{\mathbf{W}}(f) = \begin{bmatrix} \alpha_{11} e^{-2\pi^2 f^2 \sigma_{11}^2} & -\alpha_{12} e^{-2\pi^2 f^2 \sigma_{12}^2} \\ \alpha_{21} e^{-2\pi^2 f^2 \sigma_{21}^2} & -\alpha_{22} e^{-2\pi^2 f^2 \sigma_{22}^2} \end{bmatrix}.$$

Therefore, we have, with the notation of Theorem 4.9,

$$\widetilde{\mathbf{W}}^*(f)\widetilde{\mathbf{W}}(f) \overset{def}{=} \mathbf{X}(f) = \begin{bmatrix} A & C \\ C & B \end{bmatrix}.$$

It can be easily verified that

$$A = \tau_1 \left(\alpha_{11}^2 \tau_1 e^{-4\pi^2 \sigma_{11}^2 f^2} + \alpha_{21}^2 \tau_2 e^{-4\pi^2 \sigma_{21}^2 f^2} \right),$$
$$B = \tau_2 \left(\alpha_{22}^2 \tau_2 e^{-4\pi^2 \sigma_{22}^2 f^2} + \alpha_{12}^2 \tau_1 e^{-4\pi^2 \sigma_{12}^2 f^2} \right),$$

and

$$C = -\sqrt{\tau_1\tau_2} \left(\alpha_{21}\alpha_{22}\tau_2 e^{-2\pi^2(\sigma_{21}^2+\sigma_{22}^2)f^2} + \alpha_{12}\alpha_{11}\tau_1 e^{-2\pi^2(\sigma_{12}^2+\sigma_{11}^2)f^2} \right).$$

By construction the eigenvalues of the matrix $\mathbf{X}$ are positive (it is Hermitian), the largest one, $\lambda_{\max}$, being given by

$$\lambda_{\max} = \frac{1}{2}\left(A + B + \sqrt{(A-B)^2 + 4C^2} \right).$$

Introducing the parameters $A_1 = (\tau_1\alpha_{11})^2$, $A_2 = (\tau_2\alpha_{22})^2$, $r = \tau_1/\tau_2$, $x_1 = \alpha_{21}/\alpha_{11}$, and $x_2 = \alpha_{12}/\alpha_{22}$, we can rewrite A, B, and C as

$$A = A_1 \left(e^{-4\pi^2\sigma_{11}^2 f^2} + \frac{x_1^2}{r} e^{-4\pi^2\sigma_{21}^2 f^2} \right), \quad B = A_2 \left(e^{-4\pi^2\sigma_{22}^2 f^2} + r x_2^2 e^{-4\pi^2\sigma_{12}^2 f^2} \right),$$

and

$$C = -\sqrt{A_1 A_2} \left(\frac{x_1}{\sqrt{r}} e^{-2\pi^2(\sigma_{21}^2+\sigma_{22}^2)f^2} + x_2\sqrt{r} e^{-2\pi^2(\sigma_{12}^2+\sigma_{11}^2)f^2} \right).$$

The necessary and sufficient condition that the two eigenvalues are less than 1 for all f is therefore $\lambda_{\max} < 1$ or

$$c(f) \overset{def}{=} 2 - A - B - \sqrt{(A-B)^2 + 4C^2} > 0 \quad \forall f. \tag{7.1}$$

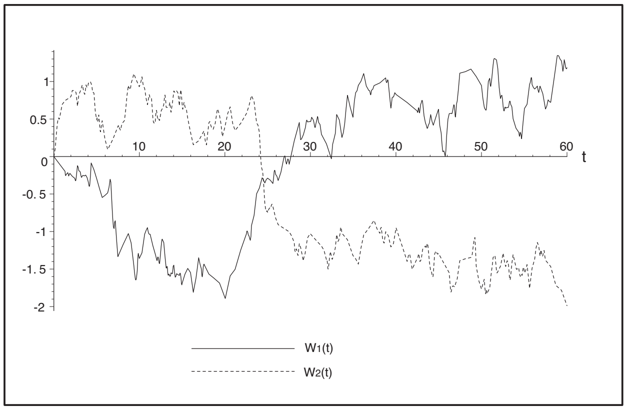


FIG. 2. *The two coordinates of the input* $\mathbf{I}_{\text{ext}}(t)$ *are realizations of independent standard Brownian/Wiener processes. Time runs along the horizontal axis.*

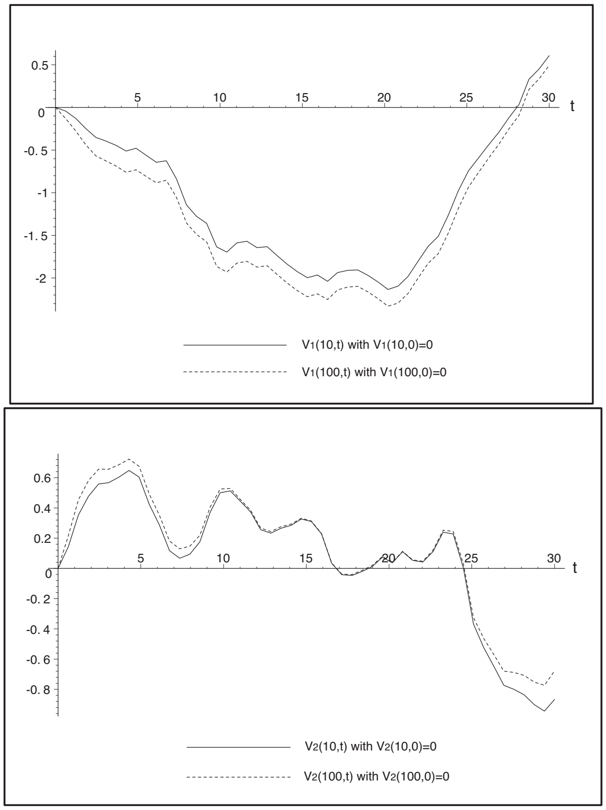


FIG. 3. *An illustration of the fact that when the connectivity matrix is translation invariant there does not exist in general a homogeneous solution: The state vectors of different neural masses follow different trajectories even when the input and the initial condition are homogeneous (independent of the location* x*). Top graph: The time variation of the first coordinate of the solution at points of coordinates* 0.1 *(continuous line) and* 1 *(dotted line) of the interval* $[0, 1]$*. Bottom graph: Same for the second coordinate. The initial condition is* 0 *in both cases.*

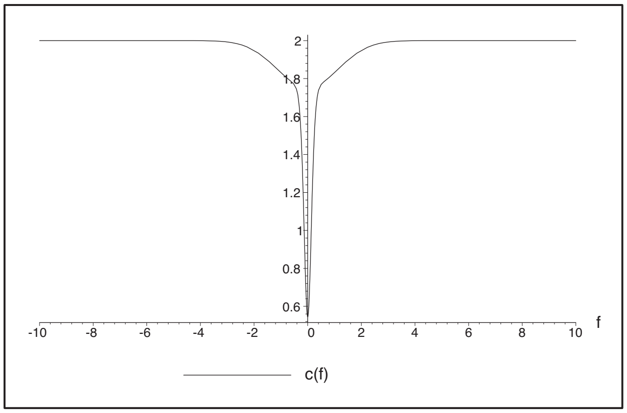


FIG. 4. *The function $c(f)$ defined in* (7.1) *is positive for all spatial frequencies f: The system is absolutely stable.*

The function $c(f)$ depends on the spatial frequency f and the nine parameters A_1, A_2, x_1, x_2, r, and $\boldsymbol{\sigma}$, the 2×2 matrix σ_{ij}, $i, j = 1, 2$.

We have solved (2.6) on $\Omega = [0, 1]$. We have sampled the interval with 100 points corresponding to 100 neural masses. The input $\mathbf{I}_{\text{ext}}$ is equal to $[W_1(t), W_2(t)]^T$, where the $W_i(t)$'s, $i = 1, 2$, are realizations of independent standard Brownian/Wiener processes shown in Figure 2. We know that the solution is not homogeneous because $\mathbf{W}$ is translation-invariant. This is illustrated in Figure 3. The initial conditions are homogeneous and equal to $(0, 0)$ for all neural mass state vectors $\mathbf{V}$.

7.1.1. Absolute stability of the solution. Let us now study the absolute stability of the solutions. According to Theorem 4.9 and the previous analysis, a sufficient condition for absolute stability is that $c(f) > 0$ for all frequencies f. As shown in Figure 4, the following choice of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ produces a curve $c(f)$ that is positive for all frequencies:

$$\boldsymbol{\alpha} = \begin{bmatrix} 2 & 1.414 \\ 1.414 & 2 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}.$$

We can check that this is indeed the case in Figure 5, which shows the absolute stability of the solution at the point of coordinate 0.5 of the interval $[0, 1]$.

7.1.2. Loss of absolute stability. The following choice of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ produces a curve $c(f)$ that is not positive for all frequencies (see Figure 6):

$$\boldsymbol{\alpha} = \begin{bmatrix} 565.7 & 565.7 \\ 565.7 & 565.7 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 0.01 & 0.01 \\ 0.1 & 0.1 \end{bmatrix}.$$

Therefore, absolute stability is not guaranteed. We show in Figure 7 that this is indeed the case.

7.2. Homogeneous solutions. In the previous case the translation invariance of the connectivity matrix forbids the existence of homogeneous solutions. We can obtain a connectivity matrix satisfying condition (3.5) by defining

$$W_{ij}(x, x') = \pm \alpha \alpha_{ij} \frac{G_{ij}(x - x')}{\int_0^1 G_{ij}(x - y)\, dy}, \quad i, j = 1, 2,$$

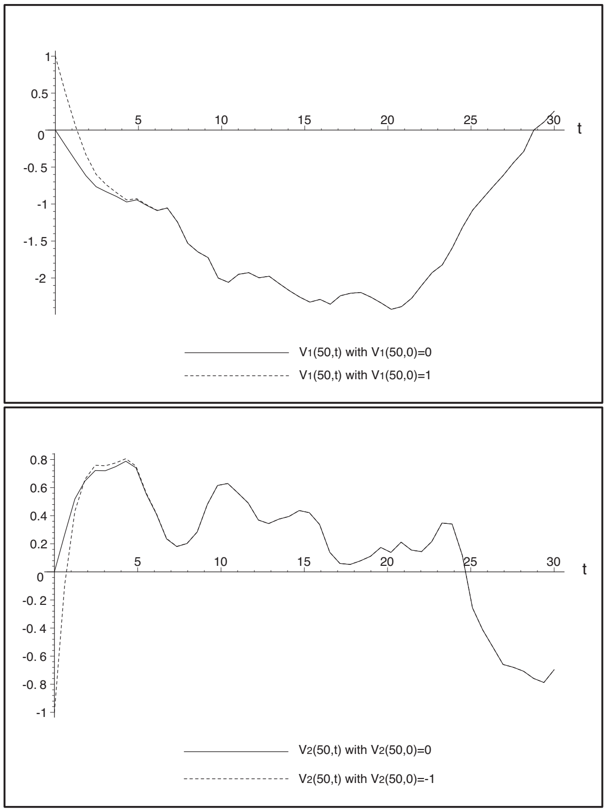


Fig. 5. *An illustration of the absolute stability of the solution: Independently of the choice of the initial condition, the trajectories of the state vector converge to a single trajectory. Results are shown for the neural mass of spatial coordinate* 0.5. *Top: The first coordinate of the state vector. Bottom: The second coordinate. Initial condition* $(0, 0)$*, continuous curves. Initial condition* $(1, -1)$*, dotted line.*

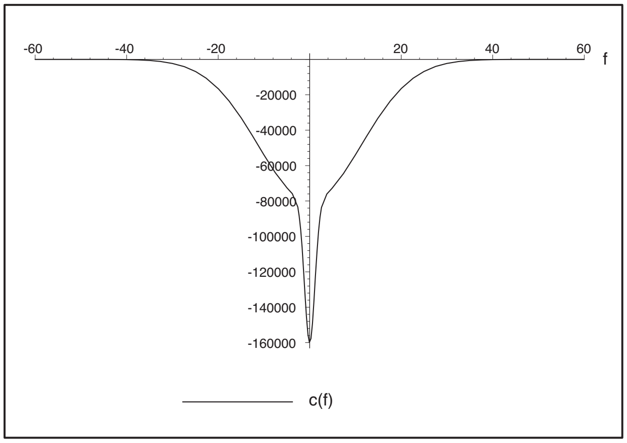


Fig. 6. *The function* $c(f)$ *defined in* (7.1) *is not positive for all spatial frequencies* f*: The system may lose its absolute stability.*

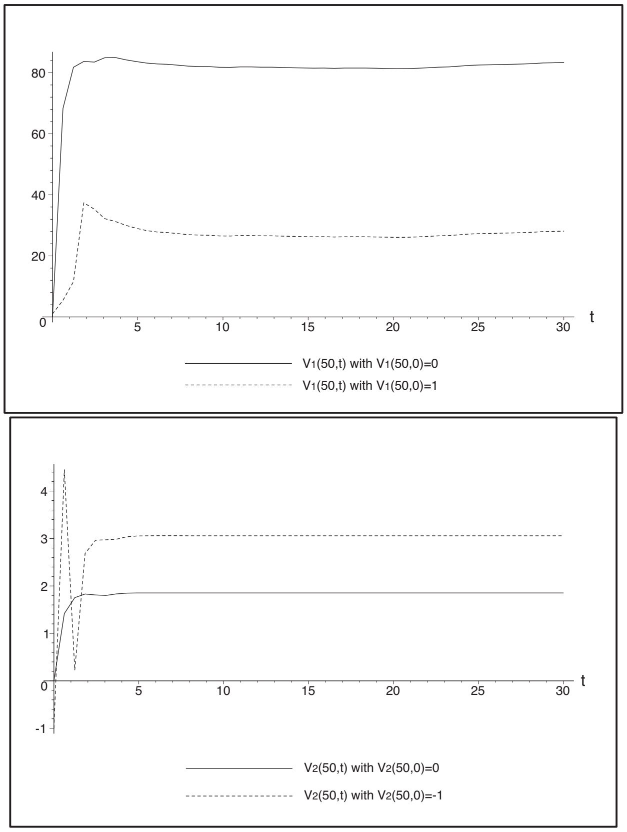


FIG. 7. *An illustration of the lack of absolute stability of the solution: Different initial conditions result in different trajectories of the state vectors. Results are shown for the neural mass of spatial coordinate* 0.5*. Top: the first coordinate of the state vector. Bottom: the second coordinate. Initial condition* $(0, 0)$*, continuous curves. Initial condition* $(1, -1)$*, dotted curves.*

where α and the α_{ij}'s are connectivity weights. These functions are well defined since the denominator is never equal to 0 and the resulting connectivity matrix is in $\mathbf{L}^2_{2\times 2}([0, 1] \times [0, 1])$. It is shown in Figure 8. The values of the parameters are given in (7.2). Proposition 6.3 guarantees the existence and uniqueness of a homogeneous solution for an initial condition in $\mathbf{L}^2_2(\Omega)$. According to Theorem 5.2 and our choice for the values of $\tau_{\max}$ and DS_m, a sufficient condition for this solution to be absolutely stable is that $\|g*\|_{\mathcal{G}_c^\perp} < 1$.

7.2.1. Absolute stability. The values of the parameters

$$\boldsymbol{\alpha} = \begin{bmatrix} 5.20 & 5.20 \\ 2.09 & 2.09 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 0.1 & 0.1 \\ 1 & 1 \end{bmatrix}, \quad \tau_1 = \tau_2 = 1, \quad \alpha = 1/20 \tag{7.2}$$

yield $\|g^*\|_{\mathcal{G}_c^\perp} \simeq 0.01$; hence the homogeneous solutions are absolutely stable. All operator norms have been computed using the method described in section A.3.

The initial conditions are drawn randomly and independently from the uniform distribution on $[-2, 2]$. The input $\mathbf{I}_{\text{ext}}(t)$ is equal to $[W_1(t), W_2(t)]^T$, where the $W_i(t)$'s, $i = 1, 2$, are realizations of independent standard Brownian/Wiener processes shown in Figure 2.

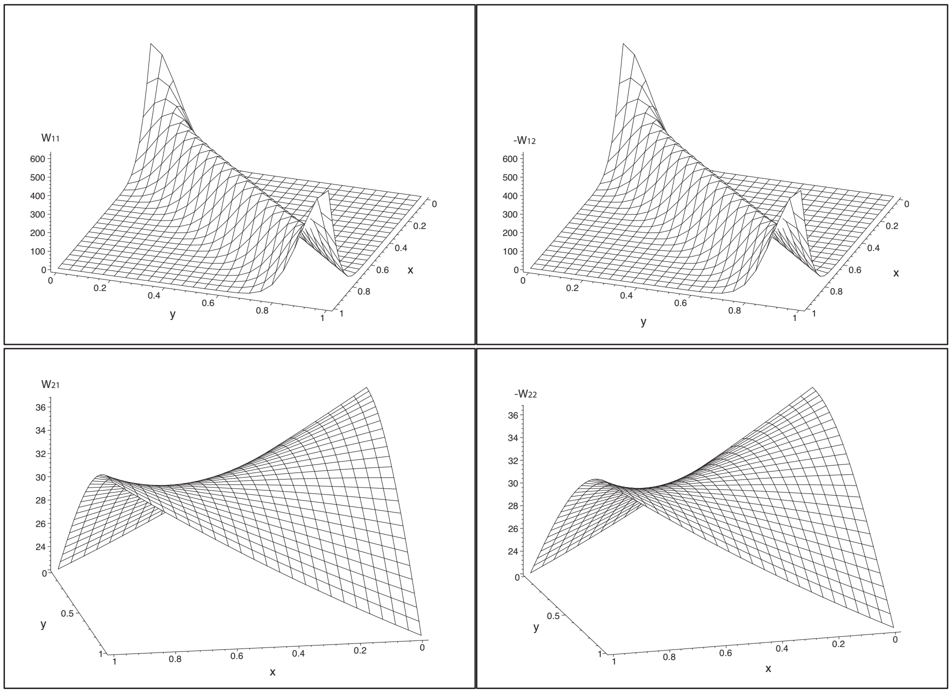


FIG. 8. *The four elements of the matrix* $\mathbf{W}(x,y)$ *in the homogeneous case. Upper left:* $W_{11}(x,y)$*. Upper right:* $-W_{12}(x,y)$*. Lower left:* $W_{21}(x,y)$*. Lower right:* $-W_{22}(x,y)$*.*

We show in Figure 9 the complete synchronization of four (numbers 10, 36, 63, and 90) of the hundred neural masses that results from the absolute stability of the homogeneous solution.

7.2.2. Loss of absolute stability. If we increase the value of α, it has the effect of increasing $\|g^*\|_{\mathcal{G}_c^\perp}$. The sufficient condition will eventually not be satisfied, and we may lose the absolute stability of the homogeneous solution and hence the complete synchronization of the solution. Such a case is shown in Figure 10 for $\alpha = 15$, corresponding to an operator norm $\|g^*\|_{\mathcal{G}_c^\perp} \simeq 2.62$.

7.3. Locally homogeneous solutions. We partition $\Omega = [0,1]$ into $\Omega_1 = [0,1/2[$ and $\Omega_2 = [1/2,1]$; hence with the notation of section 6.2, $P = 2$. We can obtain a connectivity matrix satisfying condition (6.2) by defining

$$W_{ij}(x,x') = \begin{cases} \pm\alpha\alpha_{ij}(x,x')\dfrac{G_{ij}(x-x')}{\displaystyle\int_0^{1/2} G_{ij}(x-y)\,dy}, \ x' \in \Omega_1, \\ \pm\alpha\alpha_{ij}(x,x')\dfrac{G_{ij}(x-x')}{\displaystyle\int_{1/2}^{1} G_{ij}(x-y)\,dy}, \ x' \in \Omega_2, \end{cases}$$

with $\alpha_{ij}(x,x') = \alpha_{ij}^{kl}$, $x \in \Omega_k$, $x' \in \Omega_l$, $k,l = 1,2$.

The resulting connectivity matrix is in $\mathbf{L}^2_{2\times 2}([0,1]\times[0,1])$. It is shown in Figure 11. The input $\mathbf{I}_{\rm ext}(t)$ is equal to $[W_1(t), W_2(t)]^T$ in Ω_1 and to $[W_3(t), W_4(t)]^T$ in Ω_2, where the $W_i(t)$'s, $i = 1,\dots,4$, are realizations of independent standard Brownian/Wiener

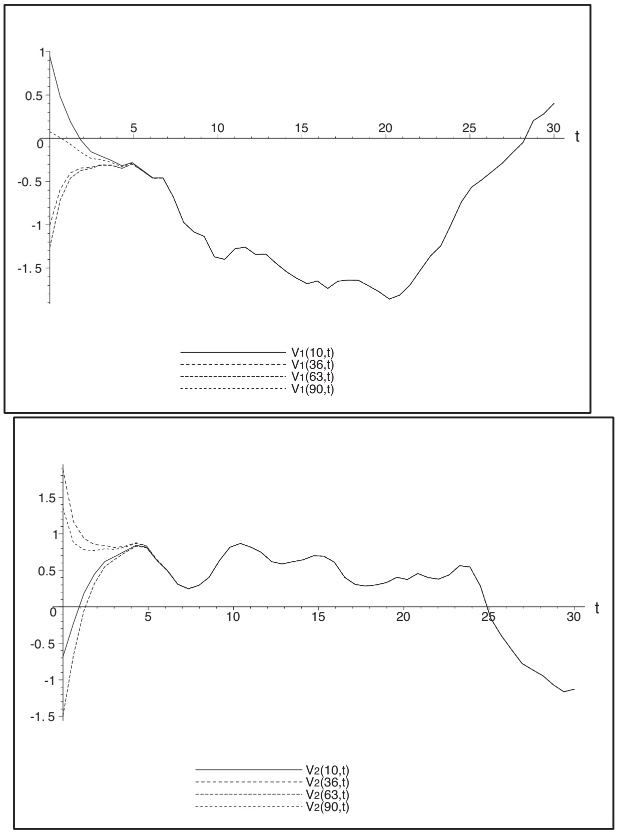


FIG. 9. *The absolute stability of the homogeneous solution results in the complete synchronization of the neural masses. This is shown for four out of the hundred (coordinates* 0.1, 0.36, 0.63, *and* 0.9*). The input is shown in Figure* 2. *The initial conditions are drawn independently from the uniform distribution on* $[-2, 2]$. *Top: The first components of the four state vectors. Bottom: The second components.*

processes shown in Figure 12. Hence it is homogeneous in Ω_1 (resp., in Ω_2) but not in $\Omega = \Omega_2 \cup \Omega_2$. According to Proposition 6.3, there exists a unique solution to (2.6) for a given initial condition in $\mathbf{L}_2^2(\Omega)$. This solution is locally homogeneous if the initial condition is locally homogeneous (Theorem 6.5), given the fact that the input is locally homogeneous.

7.3.1. Absolute stability. The parameters

$$\boldsymbol{\alpha}^{11} = \begin{bmatrix} 5.21 & 0.23 \\ 0.23 & 5.21 \end{bmatrix}, \quad \boldsymbol{\alpha}^{12} = \begin{bmatrix} 4.98 & 0.34 \\ 0.34 & 4.98 \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} 0.05 & 0.075 \\ 0.1 & 0.03 \end{bmatrix}$$
$$\boldsymbol{\alpha}^{21} = \begin{bmatrix} 4.75 & 0.45 \\ 0.45 & 4.75 \end{bmatrix}, \quad \boldsymbol{\alpha}^{22} = \begin{bmatrix} 5.39 & 0.13 \\ 0.13 & 5.39 \end{bmatrix},$$

result in an operator norm $\|g^*\|_{\mathcal{G}_c^2 \perp} \simeq 0.23$. Therefore, according to Theorem 6.7, the locally homogeneous solutions are absolutely stable, resulting in the complete local synchronization of the neural masses (within Ω_1 and Ω_2).

We show in Figure 13 (resp., Figure 14) the complete synchronization of two neural masses (numbers 10 and 36) in Ω_1 (resp., two neural masses (numbers 63

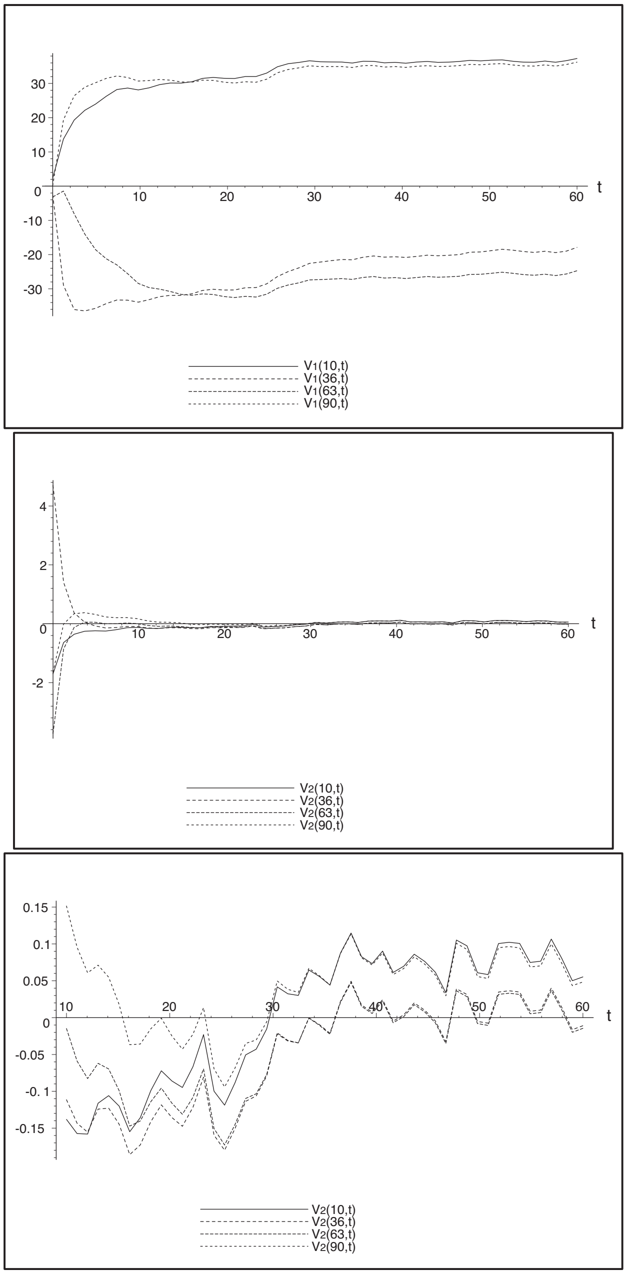


Fig. 10. *The loss of the absolute stability of the homogeneous solution results in the loss of the complete synchronization of neural masses when the sufficient condition of Theorem* 5.2 *is not satisfied. This is shown for four out of the hundred (coordinates* 0.1, 0.36, 0.63, *and* 0.9*). The input is the same as in the previous example. Top: The first components of the four state vectors. Middle: The second components of the four state vectors for* $0 \leq t \leq 60s$*. Bottom: Zoom on the second components of the four state vectors for* $10 \leq t \leq 60s$*.*

and 90) in Ω_2). The initial conditions are drawn randomly and independently from the uniform distribution on $[-10, 10]$ and $[-2, 2]$ for Ω_1 and on $[-20, 20]$ and $[-2, 2]$

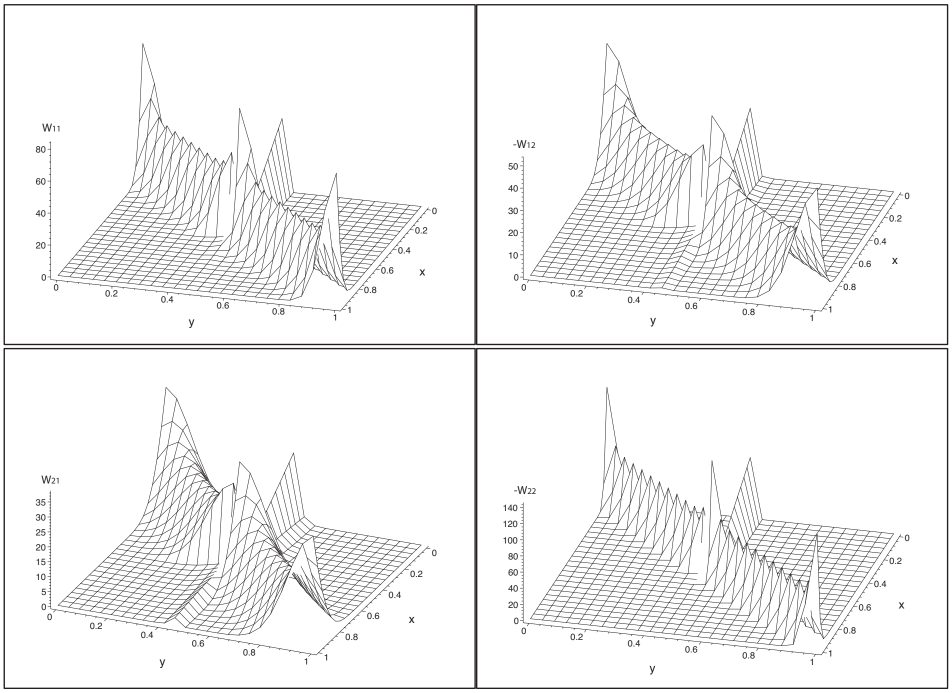


FIG. 11. *The four elements of the matrix* $\mathbf{W}(x,y)$ *in the locally homogeneous case. Upper left:* $W_{11}(x,y)$*. Upper right:* $-W_{12}(x,y)$*. Lower left:* $W_{21}(x,y)$*. Lower right:* $-W_{22}(x,y)$*.*

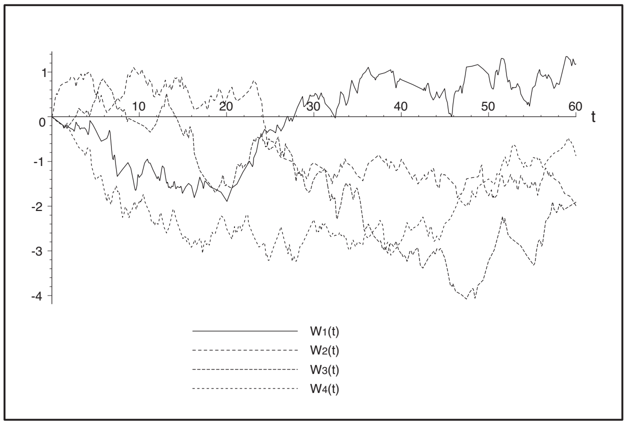


FIG. 12. *The two coordinates of the input* $\mathbf{I}_{\text{ext}}(t)$ *in* Ω_1 *and* Ω_2 *are realizations of four independent Wiener processes (*W_1 *and* W_2 *are identical to those shown in Figure* 2*).*

for Ω_2.

7.3.2. Loss of absolute stability. If we increase the value of α, it has the effect of increasing $\|g^*\|_{\mathcal{G}_c^{2},\perp}$. The sufficient condition for absolute stability will eventually not be satisfied, and we may lose the absolute stability of the locally homogeneous solution and hence the complete local synchronization of the solution. This is shown in Figures 15 and 16 for $\alpha = 10$, corresponding to an operator norm $\|g_m^{L\,*}\|_{\mathcal{G}_c^2\perp} \simeq 2.3$.

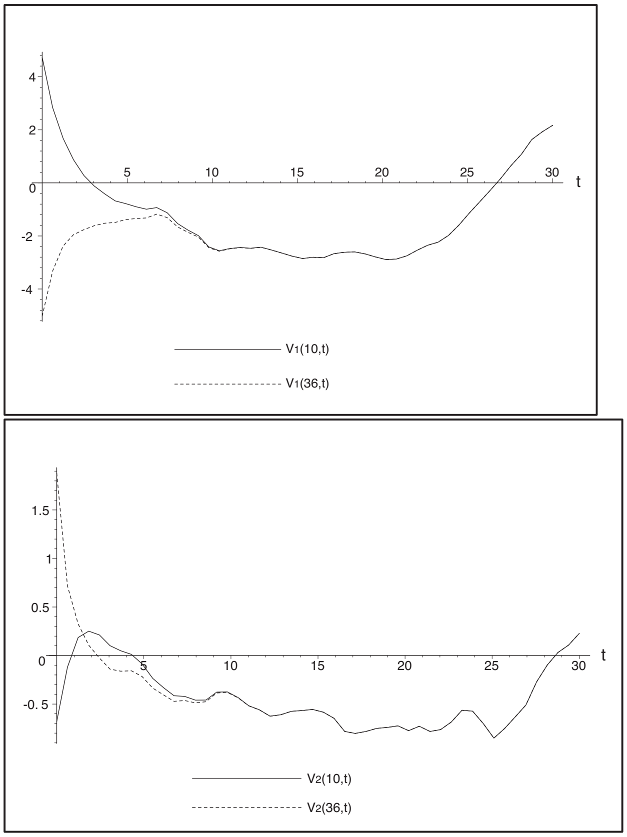


FIG. 13. *The complete synchronization of two neural masses in Ω_1 of coordinates* 0.1 *and* 0.36. *The input is shown in Figure* 12. *Top: The first components of the two state vectors. Bottom: The second components of the two state vectors.*

7.4. Pseudo–locally homogeneous solutions and their absolute stability. As mentioned at the end of section 6.2.2, even if the connectivity function does not satisfy condition (6.2) and the operator g^* satisfies only the condition of Theorem 4.7 but not that of Theorem 6.7, the existence of locally homogeneous solutions is not guaranteed, but the absolute stability of the solution is, because of Proposition 6.8. As shown in Figures 17 and 18, these solutions can be very close to being locally homogeneous and thus enjoy the property of complete local synchronization. This is potentially very interesting from the application viewpoint since one may say that if the system admits homogeneous solutions and if they are absolutely stable, it can have locally homogeneous solutions without "knowing" the partition, and they are absolutely stable.

8. Conclusion. We have studied the existence, uniqueness, and absolute stability of a solution of two examples of nonlinear integro-differential equations that describe the spatio-temporal activity of sets of neural masses. These equations involve space- and time-varying, possibly nonsymmetric, intracortical connectivity kernels. The time dependency of the connectivity kernels opens the door to the study, in this framework, of plasticity and learning. Contributions from white matter afferents are represented by external inputs. Sigmoidal nonlinearities arise from the relation

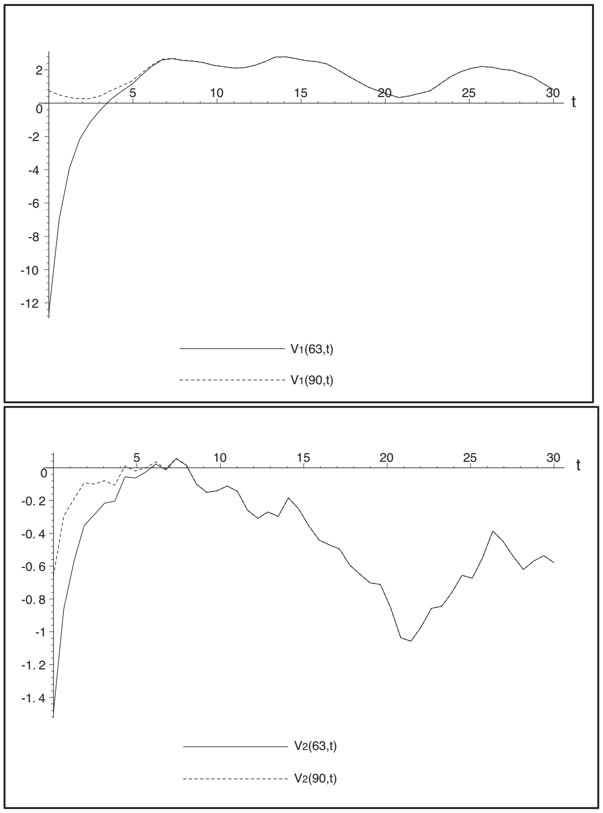


Fig. 14. *The complete synchronization of two neural masses in* Ω_2 *of coordinates* 0.63 *and* 0.9. *The input is shown in Figure* 12. *Top: The first components of the two state vectors. Bottom: The second components of the two state vectors.*

between average membrane potentials and instantaneous firing rates.

The intracortical connectivity functions have been shown to naturally define compact operators of the functional space of interest. Using methods of functional analysis, we have given sufficient conditions for the existence and uniqueness of a solution of these equations for general, homogeneous (i.e., independent of the spatial variable), and locally homogeneous inputs. In all cases we have provided sufficient conditions for the solutions to be absolutely stable, that is to say, independent of the initial state of the neural field. These conditions involve the connectivity functions and the maximum slopes of the sigmoids as well as the time constants used to described the time variation of the postsynaptic potentials. They are very relevant to neuroscience, where dynamical neuronal systems that "recognize" a given input regardless of their initial state are quite common.

To our knowledge this is the first time that such a complete analysis of the problem of the existence and uniqueness of a solution of these equations has been obtained. An important contribution also is the analysis of the absolute stability of these solutions which had been considered as much more difficult to perform than the linear stability analysis which it implies.

The reason we have been able to complete this work is our use of the functional analysis framework and the theory of compact operators in a Hilbert space with the

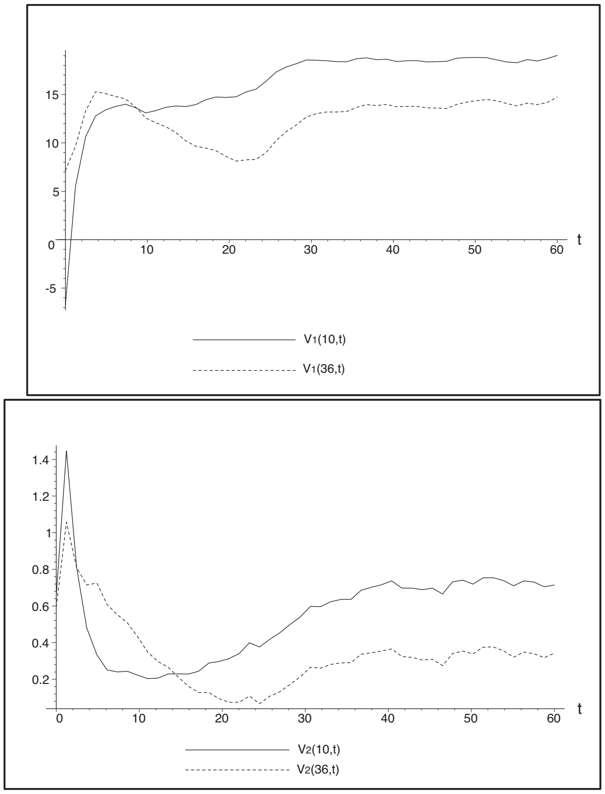


FIG. 15. *The loss of the absolute stability of the locally homogeneous solution results in the loss of the complete local synchronization of neural masses when the sufficient condition of Theorem* 6.7 *is not satisfied. This is shown for two out of the fifty (coordinates* 0.1, 0.36*) neural masses in* Ω_1*. The input is the same as in the previous example. Top: The first components of the two state vectors. Bottom: The second components of the two state vectors.*

effect of providing simple mathematical answers to some of the questions raised by modelers in neuroscience.

Future work includes adding delays to account for the distance traveled by the spikes down the axons and taking into account specific forms of the time variation of the connectivity matrices in the context of neural plasticity.

Appendix A. Notation and background material.

A.1. Matrix norms and spaces of functions. We denote $\mathcal{M}_{n\times n}$ as the set of $n \times n$ real matrices. We consider the matrix norm

$$\|\mathbf{M}\|_\infty = \max_i \sum_j |M_{ij}|.$$

We denote $\mathbf{C}_{n\times n}(\Omega)$ as the set of continuous functions from Ω to $\mathcal{M}_{n\times n}$ with the infinity norm. This is a Banach space for the norm induced by the infinity norm on $\mathcal{M}_{n\times n}$. Let $\mathbf{M}$ be an element of $\mathbf{C}_{n\times n}(\Omega)$; we denote and define $\|\mathbf{M}\|_{n\times n,\infty}$ as

$$\|\mathbf{M}\|_{n\times n,\infty} = \sup_{\mathbf{r}\in\Omega} \max_i \sum_j |M_{ij}(\mathbf{r})| = \max_i \sup_{\mathbf{r}\in\Omega} \sum_j |M_{ij}(\mathbf{r})|.$$

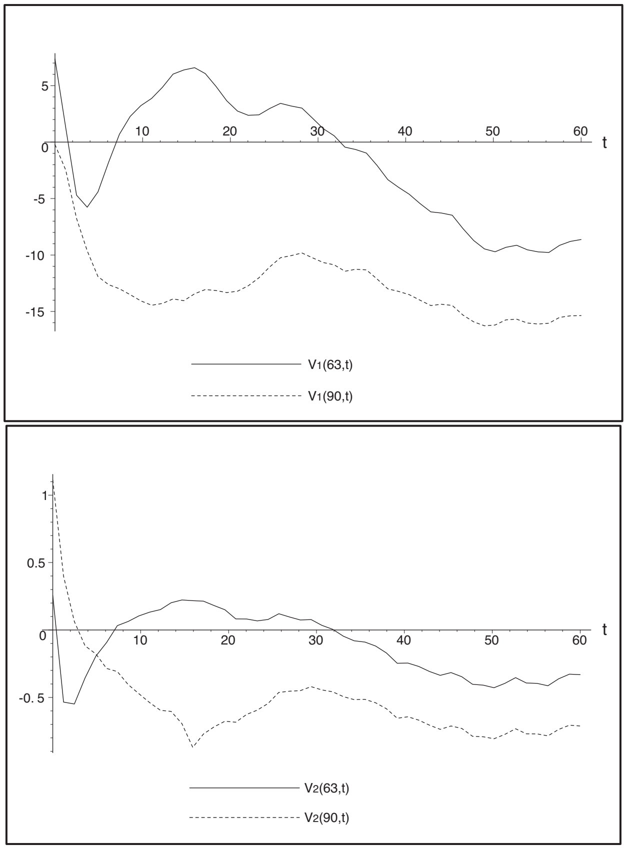


FIG. 16. *The loss of the absolute stability of the locally homogeneous solution results in the loss of the complete local synchronization of neural masses when the sufficient condition of Theorem* 6.7 *is not satisfied. This is shown for two out of the fifty (coordinates* 0.63, 0.9*) neural masses in* Ω_2. *The input is the same as in the previous example. Top: The first components of the two state vectors. Bottom: The second components of the two state vectors.*

We also denote $\mathbf{C}_n(\Omega)$ as the set of continuous functions from Ω to $\mathbb{R}^n$ with the infinity norm. This is also a Banach space for the norm induced by the infinity norm of $\mathbb{R}^n$. Let $\mathbf{x}$ be an element of $\mathbf{C}_n(\Omega)$; we denote and define $\|\mathbf{x}\|_{n,\infty}$ as

$$\|\mathbf{x}\|_{n,\infty} = \sup_{\mathbf{r}\in\Omega} \|\mathbf{x}(\mathbf{r})\|_\infty = \sup_{\mathbf{r}\in\Omega} \max_i |x_i(\mathbf{r})| = \max_i \sup_{\mathbf{r}\in\Omega} |x_i(\mathbf{r})|.$$

We can similarly define the norm $\|.\|_{n\times n,\infty}$ (resp., $\|.\|_{n,\infty}$) for the space $\mathbf{C}_{n\times n}(\Omega\times\Omega)$ (resp., $\mathbf{C}_n(\Omega\times\Omega)$).

We have the following lemma.

LEMMA A.1. *Given* $\mathbf{x}\in\mathbf{C}_n(\Omega)$ *and* $\mathbf{M}\in\mathbf{C}_{n\times n}(\Omega)$, *we have*

$$\|\mathbf{M}\,\mathbf{x}\|_{n,\infty} \le \|\mathbf{M}\|_{n\times n,\infty}\,\|\mathbf{x}\|_{n,\infty}.$$

More precisely, we have for all $\mathbf{r}\in\Omega$

$$\|\mathbf{M}(\mathbf{r})\,\mathbf{x}(\mathbf{r})\|_\infty \le \|\mathbf{M}(\mathbf{r})\|_\infty \|\mathbf{x}(\mathbf{r})\|_\infty.$$

The same results hold for $\Omega\times\Omega$ *instead of* Ω.

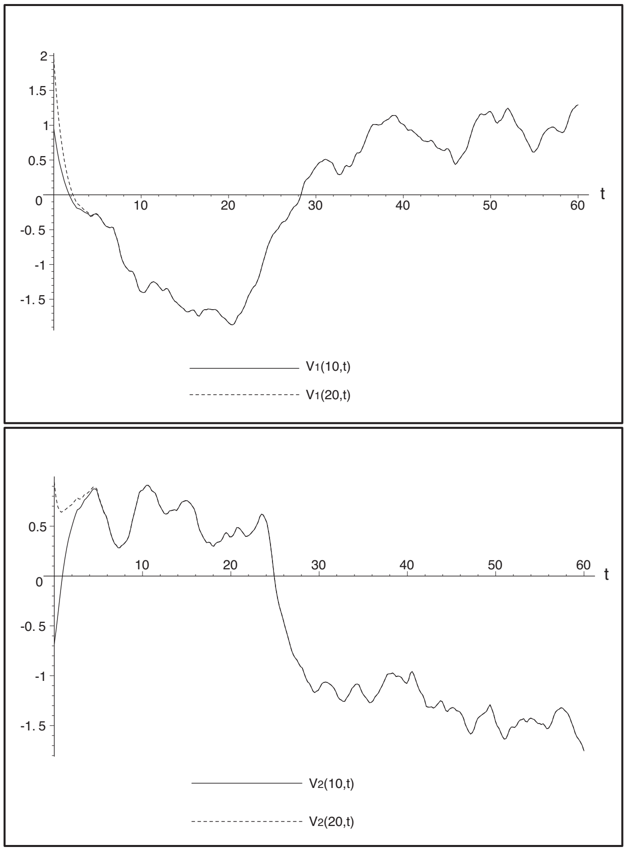


FIG. 17. *The connectivity function satisfies condition* (3.5) *but not condition* (6.2)*, and the operator g^* satisfies the condition of Theorem* 5.2 *but not that of Theorem* 6.7*. The input is locally homogeneous, as in Figures* 13 *and* 14*. The solution is absolutely stable, because of Theorem* 5.2*, and almost locally homogeneous. Something very close to complete local synchronization is observed. This is shown for two out of the fifty (coordinates* 0.1*,* 0.2*) neural masses in Ω_1. Top: The first components of the two state vectors. Bottom: The second components of the two state vectors.*

Proof. Letting $\mathbf{y} = \mathbf{M}\,\mathbf{x}$, we have

$$y_i(\mathbf{r}) = \sum_j M_{ij}(\mathbf{r}) x_j(\mathbf{r}),$$

and therefore

$$|y_i(\mathbf{r})| \leq \sum_j |M_{ij}(\mathbf{r})|\,|x_j(\mathbf{r})| \leq \sum_j |M_{ij}(\mathbf{r})|\,\|\mathbf{x}(\mathbf{r})\|_\infty,$$

so, taking the $\max_i$,

$$\|\mathbf{y}(\mathbf{r})\|_\infty \leq \|\mathbf{M}(\mathbf{r})\|_\infty\,\|\mathbf{x}(\mathbf{r})\|_\infty,$$

from which the first statement easily comes. □

We also consider the Frobenius norm on $\mathcal{M}_{n\times n}$

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i,j=1}^{n} M_{ij}^2}$$

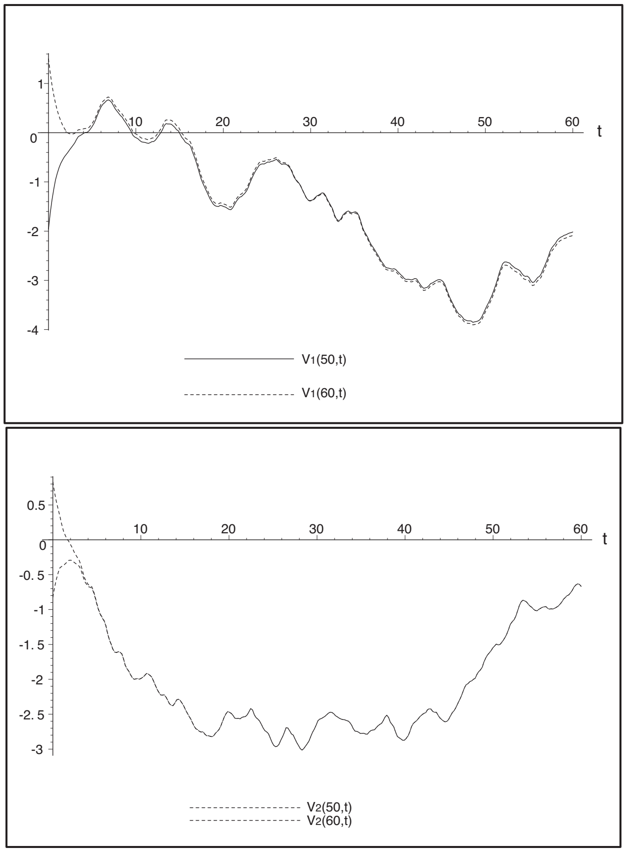


FIG. 18. *Same as in Figure* 17*. The complete local synchronization is shown for two out of the fifty (coordinates* 0.5, 0.6*) neural masses in* Ω_2.

and consider the space $\mathbf{L}^2_{n\times n}(\Omega\times\Omega)$ of the functions from $\Omega\times\Omega$ to $\mathcal{M}_{n\times n}$, whose Frobenius norm is in $L^2(\Omega\times\Omega)$. If $\mathbf{W}\in\mathbf{L}^2_{n\times n}(\Omega\times\Omega)$, we denote $\|\mathbf{W}\|^2_F = \int_{\Omega\times\Omega}\|\mathbf{W}(\mathbf{r},\mathbf{r}')\|^2_F\,d\mathbf{r}\,d\mathbf{r}'$. Note that this implies that each element w_{ij}, $i,j=1,\dots,n$ is in $L^2(\Omega\times\Omega)$. We denote $\mathbf{L}^2_n(\Omega)$ as the set of square-integrable mappings from Ω to $\mathbb{R}^n$ and $\|\mathbf{x}\|_{n,2} = (\sum_j \|x_j\|^2_2)^{1/2}$ as the corresponding norm. We have the following lemma.

LEMMA A.2. *Given* $\mathbf{x}\in\mathbf{L}^2_n(\Omega)$ *and* $\mathbf{W}\in\mathbf{L}^2_{n\times n}(\Omega\times\Omega)$*, we define* $\mathbf{y}(\mathbf{r}) = \int_\Omega \mathbf{W}(\mathbf{r},\mathbf{r}')\mathbf{x}(\mathbf{r}')\,d\mathbf{r}'$*. This integral is well defined for almost all* $\mathbf{r}$*,* $\mathbf{y}$ *is in* $\mathbf{L}^2_n(\Omega)$*, and we have*

$$\|\mathbf{y}\|_{n,2}\le\|\mathbf{W}\|_F\,\|\mathbf{x}\|_{n,2}.$$

Proof. Since each w_{ij} is in $L^2(\Omega\times\Omega)$, $w_{ij}(\mathbf{r},.)$ is in $L^2(\Omega)$ for almost all $\mathbf{r}$, thanks to Fubini's theorem. So $w_{ij}(\mathbf{r},.)x_j(.)$ is integrable for almost all $\mathbf{r}$ from our deduction that $\mathbf{y}$ is well defined for almost all $\mathbf{r}$. Next, we have

$$|y_i(\mathbf{r})|\le\sum_j\left|\int_\Omega w_{ij}(\mathbf{r},\mathbf{r}')\,x_j(\mathbf{r}')\,d\mathbf{r}'\right|$$

and (Cauchy–Schwarz)

$$|y_i(\mathbf{r})| \le \sum_j \left(\int_\Omega w_{ij}^2(\mathbf{r}, \mathbf{r}')\, d\mathbf{r}' \right)^{1/2} \|x_j\|_2,$$

from which it follows that (Cauchy–Schwarz again, discrete version)

$$\begin{aligned} |y_i(\mathbf{r})| &\le \left(\sum_j \|x_j\|_2^2 \right)^{1/2} \left(\sum_j \int_\Omega w_{ij}^2(\mathbf{r}, \mathbf{r}')\, d\mathbf{r}' \right)^{1/2} \\ &= \|\mathbf{x}\|_{n,2} \left(\sum_j \int_\Omega w_{ij}^2(\mathbf{r}, \mathbf{r}')\, d\mathbf{r}' \right)^{1/2}, \end{aligned}$$

from which it follows that $\mathbf{y}$ is in $\mathbf{L}_n^2(\Omega)$ (thanks again to Fubini's theorem) and

$$\|\mathbf{y}\|_{n,2}^2 \le \|\mathbf{x}\|_{n,2}^2 \sum_{i,j} \int_{\Omega\times\Omega} w_{ij}^2(\mathbf{r}, \mathbf{r}')\, d\mathbf{r}'\, d\mathbf{r} = \|\mathbf{x}\|_{n,2}^2 \, \|\mathbf{W}\|_F^2. \qquad \square$$

A.2. Banach space-valued functions. A useful viewpoint that is used in this article is to consider the state vector of the neural field as a mapping from a closed time interval J containing the origin 0 into one of the spaces discussed in the previous section. We denote $C(\mathrm{J}; \mathbf{C}_{\mathrm{n}}(\Omega))$ as the set of continuous mappings from J to the Banach space $\mathbf{C}_n(\Omega)$ and $C(\mathrm{J}; \mathbf{L}_{\mathrm{n}}^2(\Omega))$ as the set of continuous mappings from J to the Hilbert (hence Banach) space $\mathbf{L}_n^2(\Omega)$; see, e.g., [11].

A.3. Computation of operator norms. We give a method to compute the norms $\|g\|_{\mathcal{G}}$ and $\|g^*\|_{\mathcal{G}_c^\perp}$ for an operator g of the form

$$g(\mathbf{x})(\mathbf{r}) = \int_\Omega \mathbf{W}(\mathbf{r}, \mathbf{r}')\, \mathbf{x}(\mathbf{r}')\, d\mathbf{r}'.$$

Since $\mathcal{G}$ (resp., $\mathcal{G}_c^\perp$) is dense in the Hilbert space $\mathbf{L}^2(\Omega)$ (resp., $\mathbf{L}_0^2(\Omega)$, the subspace of $\mathbf{L}^2(\Omega)$ of functions with zero mean), we have $\|g\|_{\mathcal{G}} = \|g\|_{\mathbf{L}^2}$ and $\|g^*\|_{\mathcal{G}_c^\perp} = \|g^*\|_{\mathbf{L}_0^2}$. We consider the compact self-adjoint operators

$$G = g^* g : \mathbf{L}^2 \to \mathbf{L}^2$$

and

$$G_c^\perp = g^* \mathcal{P} g : \mathbf{L}_0^2 \to \mathbf{L}_0^2,$$

where $\mathcal{P}$ is the orthogonal projection on $\mathbf{L}_0^2$. We compute the norms of the two self-adjoint positive operators G and $G_c^\perp$ and use the relations

$$\|G\|_{\mathbf{L}^2} = \|g\|_{\mathbf{L}^2}^2$$

and

$$\|G_c^\perp\|_{\mathbf{L}_0^2} = \|g^* \mathcal{P}^* \mathcal{P} g\|_{\mathbf{L}_0^2} = \|g^* \mathcal{P}^*\|_{\mathbf{L}_0^2}^2 = \|g^*\|_{\mathbf{L}_0^2}^2.$$

Let T be a compact self-adjoint positive operator on a Hilbert space $\mathcal{H}$. Its largest eigenvalue is $\lambda = \|T\|_{\mathcal{H}}$. Let $x \in \mathcal{H}$. If $x \notin \mathrm{Ker}(\lambda \mathrm{Id} - T)^{\perp}$, then, according to, e.g., [9],

$$\lim_{n \to \infty} \|T^n x\|_{\mathcal{H}} / \|T^{n-1} x\|_{\mathcal{H}} = \lambda.$$

This method can be applied to g_m and h_m and generalized to the computation of the $\|.\|_{\mathcal{G}_c^{P\,\perp}}$ norm.

Appendix B. Global existence of solutions. In this appendix, we complete the proof of Proposition 3.4 by computing the constant $\tau > 0$ such that for any initial condition $(t_0, \mathbf{V}_0) \in \mathbb{R} \times \mathcal{F}$, the existence and uniqueness of the solution $\mathbf{V}$ are guaranteed on the closed interval $[t_0 - \tau,\, t_0 + \tau]$.

We refer the reader to [2] and exploit the following theorem.

THEOREM B.1. *Let $\mathcal{F}$ be a Banach space and $c > 0$. We consider the initial value problem*

$$\begin{cases} \mathbf{V}'(t) &= f(t, \mathbf{V}(t)), \\ \mathbf{V}(t_0) &= \mathbf{V}_0 \end{cases}$$

for $|t - t_0| < c$, where $\mathbf{V}_0$ is an element of $\mathcal{F}$ and $f : [t_0 - c,\, t_0 + c] \times \mathcal{F} \to \mathcal{F}$ is continuous. Let $b > 0$. We define the set $Q_{b,c} \equiv \{(t, \mathbf{X}) \in \mathbb{R} \times \mathcal{F},\ |t - t_0| \leq c$ and $\|\mathbf{X} - \mathbf{V}_0\| \leq b\}$. Assume the function $f : Q_{b,c} \to \mathcal{F}$ is continuous and uniformly Lipschitz continuous with respect to its second argument, i.e.,

$$\|f(t, \mathbf{X}) - f(t, \mathbf{Y})\| \leq K_{b,c} \|\mathbf{X} - \mathbf{Y}\|,$$

where $K_{b,c}$ is a constant independent of t.

Let $M_{b,c} = \sup_{Q_{b,c}} \|f(t, \mathbf{X})\|$ and $\tau_{b,c} = \min\{b/M_{b,c}, c\}$.

Then the initial value problem has a unique continuously differentiable solution $\mathbf{V}(.)$ defined on the interval $[t_0 - \tau_{b,c},\, t_0 + \tau_{b,c}]$.

In our case, $f = f_v$ and all the hypotheses of the theorem hold, thanks to Proposition 3.2 and the hypotheses of Proposition 3.4, with

$$K_{b,c} = \|\mathbf{L}\|_{\infty} + |\Omega|\, D S_m \sup_{|t - t_0| \leq c} \|\mathbf{W}(\cdot, \cdot, t)\|_{n \times n, \infty},$$

where the sup is well defined (continuous function on a compact domain).

We have

$$M_{b,c} \leq \|\mathbf{L}\|_{\infty} \left(\|\mathbf{V}_0\|_{n,\infty} + b \right) + |\Omega|\, S_m\, W + I,$$

where $W = \sup_{|t - t_0| \leq c} \|\mathbf{W}(\cdot, \cdot, t)\|_{n \times n, \infty}$ and $I = \sup_{|t - t_0| \leq c} \|\mathbf{I}_{\mathrm{ext}}(\cdot, t)\|_{n, \infty}$.

So

$$b/M_{b,c} \geq \frac{1}{\|\mathbf{L}\|_{\infty} + \frac{\|\mathbf{L}\|_{\infty} \|\mathbf{V}_0\|_{n,\infty} + |\Omega|\, S_m\, W + I}{b}}.$$

Hence, for $c \geq \frac{1}{2\|\mathbf{L}\|_{\infty}}$ and b big enough, we have $\tau_{b,c} \geq \frac{1}{2\|\mathbf{L}\|_{\infty}}$, and we can set $\tau = \frac{1}{2\|\mathbf{L}\|_{\infty}}$.

A similar proof applies to the case $f = f_a$ and that of Proposition 6.3.

Appendix C. More on $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\lambda}_{\mathbf{max}}$. Expressing the exponential as a power series in the definition of $\boldsymbol{\mathcal{M}}$ and computing the powers of the block matrix $\boldsymbol{\mathcal{L}}$, we easily find a block expression of $\boldsymbol{\mathcal{M}}$ depending on $\mathbf{L}$:

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \mathbf{L}/4 + 5\mathbf{L}^{-1}/4 & \mathbf{L}^{-2}/2 \\ \mathbf{L}^{-2}/2 & \mathbf{L}^{-1}/4 + \mathbf{L}^{-3}/4 \end{pmatrix}.$$

$\boldsymbol{\mathcal{M}}$ is diagonalizable, as a symmetric positive definite matrix, and has at most $2n$ distinct eigenvalues. More precisely, these eigenvalues are the roots of the second-order polynomials

$$\lambda^2 - \left(\frac{1}{4\,\tau_i} + \frac{3\,\tau_i}{2} + \frac{\tau_i^3}{4}\right)\lambda + \frac{1}{16} + \frac{3\,\tau_i^2}{8} + \frac{\tau_i^4}{16}, \quad 1 \le i \le n.$$

The largest eigenvalue of each of these polynomials is

$$\lambda(\tau_i) = \frac{1}{8\,\tau_i}\left(1 + 6\,\tau_i^2 + \tau_i^4 + \sqrt{1 + 8\,\tau_i^2 + 14\,\tau_i^4 + 8\,\tau_i^6 + \tau_i^8}\right),$$

so that $\lambda_{\max}$ is simply $\max_i \lambda(\tau_i)$. Note that since the function $\lambda(\tau)$ is not monotonous, $\lambda_{\max}$ is not necessarily equal to $\lambda(\tau_{\max})$.

Appendix D. Summary of important notation. Table D.1 summarizes some notation which is introduced in the article and is used in several places.

TABLE D.1
Summary of some important definitions.

Matrix functions	Definition (if applicable)	Where defined	Operators (if applicable)
$\mathbf{L}$	diagonal matrix of the inverse synaptic time constants	(2.4)	
$\tau_{\max}$	largest time constant	Definition 2.2	
$D\mathbf{S}_m$		Definition 2.1	
$\mathbf{W}$		(3.2), (3.3), (3.4)	f_v, f_a, g_v
$\mathbf{W}_{cm}$	$\mathbf{W}D\mathbf{S}_m$	Definition 4.1	g_m
$\mathbf{W}_{mc}$	$D\mathbf{S}_m\mathbf{W}$	Definition 4.1	h_m

REFERENCES

[1] F. M. Atay and A. Hutt, *Stability and bifurcations in neural fields with finite propagation speed and general connectivity*, SIAM J. Appl. Math., 65 (2005), pp. 644–666.

[2] K. Atkinson and W. Han, *Theoretical Numerical Analysis*, Springer-Verlag, New York, 2001.

[3] D. P. Buxhoeveden and M. F. Casanova, *The minicolumn hypothesis in neuroscience*, Brain, 125 (2002), pp. 935–951.

[4] L. M. Chalupa and J. S. Werner, eds., *The Visual Neurosciences*, MIT Press, Cambridge, MA, 2004.

[5] T. P. Chen and S. I. Amari, *Exponential convergence of delayed dynamical systems*, Neural Computation, 13 (2001), pp. 621–635.

[6] S. Coombes, *Waves, bumps, and patterns in neural fields theories*, Biol. Cybernet., 93 (2005), pp. 91–108.

[7] S. Coombes, N. A. Venkov, L. Shiau, I. Bojak, D. T. J. Liley, and C. R. Laing, *Modeling electrocortical activity through local approximations of integral neural field equations*, Phys. Rev. E (3), 76 (2007), 51901.

[8] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, Cambridge, MA, 2001.

[9] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[10] B. Ermentrout, *Neural networks as spatio-temporal pattern-forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[11] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[12] Y. Fang and T. G. Kincaid, *Stability analysis of dynamical neural networks*, IEEE Trans. Neural Networks, 7 (1996), pp. 996–1006.

[13] O. Faugeras and F. Grimbert, *Bumps and waves in a two-dimensional multilayer neural field model*, in Proceedings of the Sixteenth Annual Computational Neuroscience Meeting (CNS) (Toronto, 2007), BMC Neuroscience 8, BioMed Central, London, 2007, pp. 49–50.

[14] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 378–407.

[15] W. J. Freeman, *Mass Action in the Nervous System*, Academic Press, New York, 1975.

[16] W. Gerstner and W. M. Kistler, *Mathematical formulations of hebbian learning*, Biol. Cybernet., 87 (2002), pp. 404–415.

[17] F. Grimbert, *Mesoscopic Models of Cortical Structures*, Ph.D. thesis, University of Nice Sophia-Antipolis, Sophia-Antipolis, France, 2008.

[18] F. Grimbert and O. Faugeras, *Bifurcation analysis of Jansen's neural mass model*, Neural Comput., 18 (2006), pp. 3052–3068.

[19] S. Haeusler and W. Maass, *A statistical analysis of information-processing properties of lamina-specific cortical microcircuits models*, Cerebral Cortex, 17 (2007), pp. 149–162.

[20] J. J. Hopfield, *Neurons with graded response have collective computational properties like those of two-state neurons*, Proc. Nat. Acad. Sci., USA, 81 (1984), pp. 3088–3092.

[21] B. H. Jansen and V. G. Rit, *Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns*, Biol. Cybernet., 73 (1995), pp. 357–366.

[22] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, *Principles of Neural Science*, 4th ed., McGraw-Hill, New York, 2000.

[23] S. Kubota and K. Aihara, *Analyzing global dynamics of a neural field model*, Neural Processing Letters, 21 (2005), pp. 133–141.

[24] C. L. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math., 63 (2002), pp. 62–97.

[25] F. H. Lopes da Silva, A. Hoeks, and L. H. Zetterberg, *Model of brain rhythmic activity*, Kybernetik, 15 (1974), pp. 27–37.

[26] F. H. Lopes da Silva, A. van Rotterdam, P. Barts, E. van Heusden, and W. Burr, *Model of neuronal populations. The basic mechanism of rhythmicity*, in Progress in Brain Research 45, M. A. Corner and D. F. Swaab, eds., Elsevier, Amsterdam, 1976, pp. 281–308.

[27] A. M. Lyapunov, *Stability of Motion*, Academic Press, New York, 1966.

[28] K. Matsuoka, *Stability conditions for nonlinear continuous neural networks with asymmetric connection weights*, Neural Networks, 5 (1992), pp. 495–500.

[29] V. B. Mountcastle, *Modality and topographic properties of single neurons of cat's somatosensory cortex*, J. Neurophysiology, 20 (1957), pp. 408–434.

[30] V. B. Mountcastle, *The columnar organization of the neocortex*, Brain, 120 (1997), pp. 701–722.

[31] Q. C. Pham and J. J. E. Slotine, *Stable concurrent synchronization in dynamic system networks*, Neural Networks, 20 (2007), pp. 62–77.

[32] A. Pikovsky, J. Kurths, and M. Rosenblum, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2001.

[33] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* 1. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.

[34] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* 2. *Lateral inhibition and standing pulses*, SIAM J. Appl. Math., 62 (2001), pp. 226–243.

[35] D. Pinto, R. Jackson, and C. E. Wayne, *Existence and stability of traveling pulses in a continuous neuronal network*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 954–984.

[36] J. J. E. Slotine and W. Li, *Applied Nonlinear Control*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[37] A. M. Thomson and A. P. Bannister, *Interlaminar connections in the neocortex*, Cerebral Cortex, 13 (2003), pp. 5–14.

[38] A. van Rotterdam, F. H. Lopes da Silva, J. van den Ende, M. A. Viergever, and A. J. Hermans, *A model of the spatial-temporal characteristics of the alpha rhythm*, Bull. Math. Biol., 44 (1982), pp. 283–305.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 251–272

TURING PATTERN FORMATION IN THE BRUSSELATOR MODEL WITH SUPERDIFFUSION[*]

A. A. GOLOVIN[†], B. J. MATKOWSKY[†], AND V. A. VOLPERT[‡]

Abstract. The effect of superdiffusion on pattern formation and pattern selection in the Brusselator model is studied. Our linear stability analysis shows, in particular, that, unlike the case of normal diffusion, the Turing instability can occur even when diffusion of the inhibitor is slower than that of the initiator. A weakly nonlinear analysis yields a system of amplitude equations, analysis of which predicts parameter regimes where hexagons, stripes, and their coexistence are expected. Numerical computations of the original Brusselator model near the stability boundaries confirm the results of the analysis. In addition, further from the stability boundaries, we find a regime of self-replicating spots.

Key words. pattern formation, Brusselator, anomalous diffusion, superdiffusion, Turing instability

AMS subject classifications. 35K57, 35Q99, 37G99

DOI. 10.1137/070703454

1. Introduction.

Reaction-diffusion systems are ubiquitous in many branches of science and engineering and have attracted the attention of scientists, engineers, and mathematicians for decades; see, e.g., [4, 9]. Since the groundbreaking discoveries of Turing [49], who showed that diffusion in a mixture of chemically reacting species could cause instability of a spatially uniform state leading to the formation of spatial patterns, and Belousov and Zhabotinskii [61, 62], who discovered oscillating chemical reactions, reaction-diffusion systems have become paradigms for spatio-temporal pattern formation in systems far from thermodynamic equilibrium [11, 31, 54], including living organisms [33]. Experimental observations of such fascinating structures as spiral waves [58], spatially regular, stationary patterns with different symmetries (hexagonal, striped, etc.) [21, 22, 39], as well as the theoretical description of chemical turbulence [19], have made reaction-diffusion systems the subject of numerous ongoing investigations.

A characteristic feature of most of the reaction-diffusion systems that have been studied to date is that diffusion is normal, i.e., at the molecular level it is the result of independent random jumps, e.g., nearest neighbor jumps, at regularly spaced time increments. In fact, the molecules can wait between successive jumps and can also execute not just nearest neighbor jumps, but rather long jumps. However, both the waiting time distribution and jump size distribution must have finite moments. In some cases, however, these conditions are not met, in that the molecules may undergo *anomalous diffusion* [3, 13, 16, 29, 30, 45]. Unlike normal diffusion, which is characterized by the dependence of the mean square displacement of a randomly walking particle on time $\langle(\Delta\mathbf{r})^2\rangle \sim t$, anomalous diffusion is characterized by the

[*] Received by the editors September 21, 2007; accepted for publication (in revised form) June 11, 2008; published electronically October 13, 2008. This work has been supported by NSF grants DMS-0707445 and DMS-0505878.

http://www.siam.org/journals/siap/69-1/70345.html

[†] Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208-3125 (a-golovin@northwestern.edu, b-matkowsky@northwestern.edu, v-volpert@northwestern.edu).

more general dependence

$$\langle(\Delta\mathbf{r})^2\rangle = 2dK_\gamma t^\gamma, \tag{1}$$

where d is the (embedding) spatial dimension, K_γ is a generalized diffusion constant, and the exponent γ is not necessarily an integer. For $\gamma = 1$, anomalous diffusion reduces to normal diffusion, with $K_1 = D$ being the ordinary diffusion coefficient. For $\gamma < 1$ ($\gamma > 1$), the diffusion process is slower (faster) than normal diffusion and is called "subdiffusive" (resp., "superdiffusive"). Both types of anomalous diffusion processes have been recognized as playing important roles in various physical, chemical, biological, and geological processes. For example, subdiffusion, which corresponds to molecules waiting for long times before jumping, i.e., with a waiting time distribution having infinite moments, often occurs in gels (especially bio-gels [53, 55]), porous media [8], and polymers [1]. An important limiting case of superdiffusion corresponds to *Lévy flights* [30], which occur in systems in which there are *long jumps* of particles, i.e., with a jump size distribution having infinite moments. It is typical of some processes in plasmas and turbulent flows [6, 12, 47], surface diffusion [25, 42], diffusion in porous media with flow [8], surfactant diffusion along polymer chains [2], cosmic ray propagation [57], charge carrier transfer in semiconductors [43], motion of animals [44, 48, 52], as well as in geophysical and geological processes, including the dispersion of nuclear waste in soil [37] (see also [3, 13, 16, 29, 30, 45] for reviews and numerous other examples).

An important property of Lévy flights is that in the continuum limit, in one dimension, they are described by means of a partial differential equation with a *fractional derivative*, which is, in fact, defined as a nonlocal, integrodifferential equation,

$$u_t = K^\mu \frac{\partial^\mu}{\partial|x|^\mu}u, \tag{2}$$

where (for $1 < \mu < 2$)

$$\frac{d^\mu u}{d|x|^\mu} = -\frac{1}{2\cos(\pi\mu/2)}(\mathbf{D}_+^\mu u + \mathbf{D}_-^\mu u), \tag{3a}$$

$$\mathbf{D}_+^\mu u = \frac{1}{\Gamma(2-\mu)}\frac{d^2}{dx^2}\int_{-\infty}^{x}\frac{u(\xi,t)\,d\xi}{(x-\xi)^{\mu-1}}, \tag{3b}$$

$$\mathbf{D}_-^\mu u = \frac{1}{\Gamma(2-\mu)}\frac{d^2}{dx^2}\int_{x}^{\infty}\frac{u(\xi,t)\,d\xi}{(\xi-x)^{\mu-1}}, \tag{3c}$$

or in a form defined by its action in Fourier space, $\mathcal{F}[\partial^\mu u/\partial|x|^\mu](k) = -|k|^\mu\mathcal{F}[u](k)$. For $\mu = 1$, (3) reduces to the derivative of the Hilbert transform of u. In higher dimensions, the Laplacian is replaced by the operator $\nabla^\beta \equiv -(-\nabla^2)^{\beta/2}(1 < \beta < 2)$, defined by its action in Fourier space, $\mathcal{F}[\nabla^\beta u](\mathbf{k}) = -|\mathbf{k}|^\beta\mathcal{F}[u](\mathbf{k})$. The derivation of this macroscopic description from an appropriate continuous time random walk model at the microscopic (i.e., molecular) level can be found in [29].

Although many aspects of anomalous diffusion have been extensively studied (see [30] for the most recent review), nonlinear dynamic and pattern formation aspects were the subject of only a very limited number of works. The propagation of reaction-diffusion fronts governed by anomalous diffusion and nonlinear kinetics (similar to the Kolmogorov–Petrovsky–Piskunov problem [18]) was investigated both for the subdiffusive case [27, 59] and the superdiffusive case [7, 26]. It was shown that superdiffusion leads to a significant increase of the front speed. It was also shown that the effects of fluctuations could be important [5].

Several papers study pattern formation in reaction-diffusion systems with anomalous diffusion [10, 14, 15, 20, 56]. The formation of Turing patterns in the subdiffusive case was studied in [56], where it was shown that subdiffusion suppresses pattern formation. In [23] it was also shown that, in one-dimensional systems, anomalous diffusion leads to anomalous heat conduction that may be very important for reaction-diffusion dynamics coupled to heat release, as in combustion. In [14,15], linear stability analysis that predicts Turing instabilities is performed, and the results are supported by numerical simulations. The nonlinear dynamics of oscillating reaction-diffusion patterns that can lead to the formation of spiral waves and chemical turbulence in systems with Levy flights has been recently studied in [34]. At the same time, important aspects of nonlinear dynamics and pattern formation in reaction-diffusion systems, such as *pattern selection* in the presence of anomalous diffusion, have not been studied.

It is well known that the diffusivities in systems of reaction-diffusion equations with normal diffusion play a decisive role in the development of instabilities leading to spatio-temporal pattern formation. For example, this is the case if the diffusivities differ significantly from one another. How much more so is this expected to be the case when one or more of the diffusivities is anomalous, where the diffusivities differ not only quantitatively, but qualitatively as well. In experiments, an efficient way to control the nature of diffusion and the dynamical regimes of chemical reactions in a liquid phase, including those generated by instabilities, is by stirring [17, 24, 28, 35, 36, 38, 41, 51, 63]. In catalytic systems with surface chemical reactions, the anomaly of the reactant surface diffusion can be controlled, for example, by turbulent flow in the adjacent gas phase.

The Brusselator is a well-known paradigm for the study of nonlinear reaction-diffusion systems. In this paper we study Turing pattern formation in the Brusselator model with superdiffusion, namely, Lévy flights. We focus on pattern selection in the formation of hexagons and stripes and compare the cases of normal and superdiffusion. In section 2 we formulate the mathematical model that we consider. It exhibits a solution corresponding to a homogeneous, stationary state. In section 3 we consider the linear stability of this state, while in section 4 we present a weakly nonlinear analysis to derive a set of coupled amplitude equations. These equations are analyzed in section 5, where we describe hexagonal and striped patterns and their stability. In section 6 we present the results of numerical computations of the original model. In addition to verifying the results of our analysis by computing hexagonal and striped patterns near the neutral stability boundaries, we also present new patterns, not described by our analysis, corresponding to the regions where hexagons and stripes are unstable. Specifically, we compute solutions corresponding to self-replicating spots. Finally, in section 7 we present conclusions of our analysis.

2. Mathematical model. The mathematical model that we consider is the standard Brusselator model (see, e.g., [50] and the references therein),

$$\frac{\partial X}{\partial \tau} = D_X \nabla^\alpha X + A - (B+1)X + X^2 Y, \tag{4a}$$

$$\frac{\partial Y}{\partial \tau} = D_Y \nabla^\beta Y + BX - X^2 Y, \tag{4b}$$

with the only exception that the Laplacians with respect to the spatial variables are replaced by ∇^α and ∇^β, the operators that represent superdiffusion as discussed earlier. The system is considered in the entire plane $-\infty < \xi, \zeta < \infty$, and we are interested in bounded solutions. We do not impose initial conditions in the analysis, as we are interested in stable steady solutions of the problem rather than in transient behaviors. We remark that the reaction terms in the equations are taken to be "normal," i.e., the same as in the normal diffusion Brusselator model. This does not have to be the case—anomalous diffusion may affect the form of the reaction terms [20, 46, 59, 60].

The critical point of the system (4), corresponding to a homogeneous stationary solution, is

$$X = A, \quad Y = \frac{B}{A}.$$

It is convenient to rewrite the system of equations in terms of the deviation of the solution from the critical point by introducing

$$U = X - A, \quad V = Y - \frac{B}{A},$$

which yields

$$\frac{\partial U}{\partial \tau} = D_X \nabla^\alpha U + (B-1)U + A^2 V + \frac{B}{A}U^2 + 2AUV + U^2 V, \tag{5a}$$

$$\frac{\partial V}{\partial \tau} = D_Y \nabla^\beta V - BU - A^2 V - \frac{B}{A}U^2 - 2AUV - U^2 V. \tag{5b}$$

Finally, we rescale (5), using

$$U = u_* u, \quad V = v_* v, \quad \tau = t_* t, \quad \xi = \ell_* x, \quad \zeta = \ell_* y$$

with

$$u_* = \sqrt{D_Y/(D_X)^{\beta/\alpha}}, \quad v_* = 1/u_*, \quad \ell_* = D_X^{1/\alpha}, \quad t_* = 1$$

to obtain

$$\frac{\partial u}{\partial t} = \nabla^\alpha u + (B-1)u + Q^2 v + \frac{B}{Q}u^2 + 2Quv + u^2 v, \tag{6a}$$

$$\eta^2 \frac{\partial v}{\partial t} = \nabla^\beta v - Bu - Q^2 v - \frac{B}{Q}u^2 - 2Quv - u^2 v, \tag{6b}$$

where

$$\eta = \sqrt{(D_X)^{\beta/\alpha}/D_Y}, \quad Q = A\eta. \tag{7}$$

The critical point is now given by $u = v = 0$.

3. Linear stability analysis. To study the linear stability of the solution $u = v = 0$, we substitute the normal mode solution

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} e^{\sigma t + ik_x x + ik_y y},$$

where σ is the growth rate of the perturbation and (k_x, k_y) is its wave vector, into (6) to obtain the dispersion relation

$$\eta^2\sigma^2 + M_1\sigma + M_2 = 0, \tag{8}$$

$$M_1 = Q^2 + k^\beta - \eta^2(B - 1 - k^\alpha), \quad M_2 = BQ^2 - (B - 1 - k^\alpha)(Q^2 + k^\beta),$$

where k is the wavenumber of the perturbation, $k = (k_x^2 + k_y^2)^{1/2}$. The problem exhibits both Turing and oscillatory instabilities. We are particularly interested in the Turing stability boundary, which corresponds to $\sigma = 0$. The neutral stability curve, which can be written in the form

$$B = \frac{1}{k^\beta}(1 + k^\alpha)(Q^2 + k^\beta),$$

has a single minimum (k_{cr}, B_{cr}) which depends on the value of Q. It can be found parametrically as

$$B_{cr} = \frac{(1+x)^2}{1 + (1-s)x}, \quad Q^2 = \frac{sx^{1+1/s}}{1 + (1-s)x}, \quad k_{cr} = x^{1/\alpha}, \tag{9}$$

where $s = \alpha/\beta$. The quantity s varies between $1/2$ and 2 since both α and β are between 1 and 2. The range of variation of the parameter x depends on s:

$$0 < x < \infty \;\text{ if }\; \frac{1}{2} < s \le 1, \quad 0 < x < \frac{1}{s-1} \;\text{ if }\; 1 < s < 2,$$

so that B_{cr} and Q^2 are positive in (9). Figure 1 demonstrates stability regions in the (Q^2, B) plane for selected values of the parameter s that are parametrically given by (9). Instability regions are located above the respective curves. We see that, except for very small Q, the general trend is that increasing s broadens the stability regions. In the weakly nonlinear analysis of the next section we will need the eigenvector (a, b) for $\sigma = 0$, $k = k_{cr}$, and $B = B_{cr}$. It is given by

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x^{1/s} \\ -(1+x) \end{pmatrix}.$$

The trivial steady state is unstable for $B > B_{cr}$. Whether it is stable for $B < B_{cr}$ depends on the location of the oscillatory instability boundary. This boundary is obtained by setting the coefficient M_1 in the dispersion relation (8) equal to zero under the condition that the coefficient M_2 is positive. Thus, oscillatory instability occurs at

$$B = 1 + \frac{Q^2}{\eta^2} \;\text{ if }\; 1 + \frac{Q^2}{\eta^2} < B_{cr}.$$

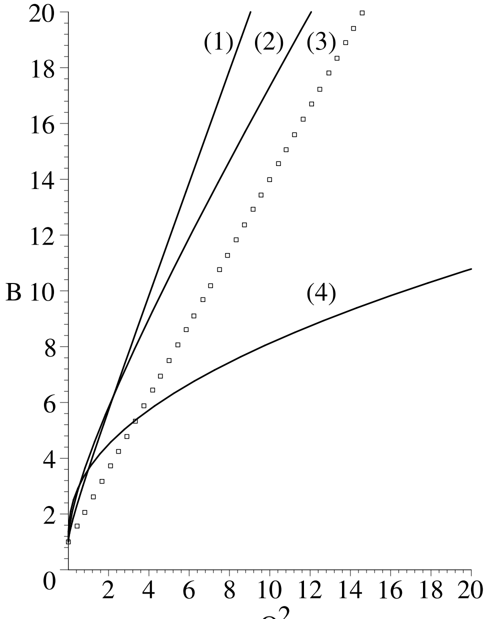


FIG. 1. *Stability boundaries in the* (Q^2, B) *plane. The instability region lies above the curve. Curve* (1) *Turing stability boundary for* $s = 2$*; curve* (2) *Turing stability boundary for* $s = 1$*; curve* (3) *oscillatory stability boundary,* $\eta = 0.877$*; curve* (4) *Turing stability boundary for* $s = 0.5$.

In order for Turing instability to occur prior to oscillatory instability as B increases, we need $B_{cr} < 1 + Q^2/\eta^2$, which can be interpreted as a condition on η,

$$\eta^2 < \frac{Q^2}{B_{cr} - 1} = \frac{s x^{1/s}}{x + 1 + s}. \tag{10}$$

The Turing mechanism of pattern formation, which is due to diffusion-induced instability of the homogeneous steady state, requires that the ratio of the diffusion coefficient of the inhibitor (in our problem, the inhibitor is Y) to that of the activator (X) be sufficiently large in "normal" reaction-diffusion systems. This corresponds to D_Y being sufficiently greater than D_X or, equivalently, η being sufficiently small. Based on this result obtained for the Brusselator problem with normal diffusion, one might expect that if the fractional derivative order β for the inhibitor Y is larger than α, the fractional derivative order for the activator X, i.e., if the inhibitor diffuses slower than the activator, $s < 1$, then no Turing instability will be observed. As can be seen from the above results, this is not necessarily true. Indeed, let us first discuss in greater detail the case of normal diffusion, in which $s = 1$. It is convenient to return to the original parameters A and B instead of Q and B. From (10), using (7) and (9), we get

$$A\eta^2 + 2\eta - A < 0,$$

so that

$$\eta = \sqrt{\frac{D_X}{D_Y}} < \sqrt{\frac{1}{A^2} + 1} - \frac{1}{A} \equiv F_1(A). \tag{11}$$

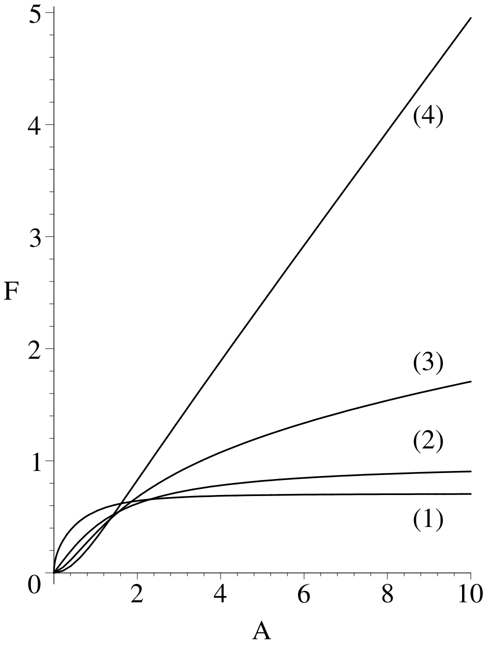


FIG. 2. *The function* $F(A)$ *for selected values of* s. *Curve* (1) $s = 2$; *curve* (2) $s = 1$; *curve* (3) $s = 0.7$; *curve* (4) $s = 0.5$.

The function $F_1(A)$ is a monotonically increasing function that is zero at $A = 0$ and goes to one as A goes to infinity. This means that if η is greater than one, i.e., the diffusion coefficient of the inhibitor is too small, then Turing instability of the homogeneous stationary solution cannot occur because it is always preceded by the oscillatory instability. If $\eta < 1$, then for the supply rate A sufficiently large, specifically, for $A > 2\eta/(1-\eta^2)$, Turing instability does occur as B increases.

In the general case of s not necessarily equal to one, condition (11) is replaced by

$$\eta^2 = \frac{D_X^{1/s}}{D_Y} < 2^{1-1/s} s \frac{[\sqrt{[A^2(s-1)+(s+1)]^2+4A^2} - A^2(s-1) - (s+1)]^{1/s}}{\sqrt{[A^2(s-1)+(s+1)]^2+4A^2} - A^2(s-1) + (s+1)} \equiv F(A), \tag{12}$$

the derivation of which also uses (7), (9), and (10). If this condition is satisfied, then by increasing B we reach the Turing instability boundary. If this condition is not satisfied, then increasing B results in oscillatory instability. We remark that increasing B necessarily results in an instability because the term without σ in the dispersion relation (8) becomes negative for B sufficiently large, so that the dispersion relation has a positive root. The function $F(A)$ in (12) is a monotonically increasing function of A with $F(0) = 0$ for any s of interest, i.e., for $1/2 \leq s \leq 2$. The behavior of this function is, however, different for $1/2 \leq s < 1$ and $1 \leq s \leq 2$. In the first case the function increases without bound as $A \to \infty$, while in the second case it remains finite with

$$\lim_{A\to\infty} F(A) = \frac{1}{s}(s-1)^{1-1/s}.$$

The behavior of this function is illustrated in Figure 2. These results are somewhat counterintuitive because not only can Turing instability occur in the case $s < 1$, when it is not expected because diffusion of the inhibitor in this case is slower than diffusion of the initiator, but it may also occur for any value of η provided that A is sufficiently large, while the case of normal diffusion carries the restriction $\eta < 1$.

4. Weakly nonlinear analysis: Stripes and hexagons. We now perform a weakly nonlinear analysis of the system (6) near the instability threshold in order to study pattern formation. Specifically, we are interested in the formation of hexagons and stripes.

We introduce the slow time $T = \varepsilon^2 t$ and expand both unknowns u and v as well as the bifurcation parameter B as

$$\begin{aligned} u &\sim \varepsilon u_1 + \varepsilon^2 u_2 + \varepsilon^3 u_3 + \cdots , \\ v &\sim \varepsilon v_1 + \varepsilon^2 v_2 + \varepsilon^3 v_3 + \cdots , \\ B &= B_{cr} + \varepsilon^2 \mu. \end{aligned} \tag{13}$$

Here u_j and v_j $(j = 1, 2, 3)$, which correspond to the long time solution behavior after all transients have decayed, are functions of T, x, and y.

Substituting the expansions (13) into the system of equations (6) and collecting like powers of ε, we obtain at orders ε^j $(j = 1, 2, 3)$ the sequence of problems

$$\begin{aligned} O(\varepsilon):\quad & L_u(u_1, v_1) \equiv \nabla^\alpha u_1 + (B_{cr} - 1)u_1 + Q^2 v_1 = 0, \\ & L_v(u_1, v_1) \equiv \nabla^\beta v_1 - B_{cr} u_1 - Q^2 v_1 = 0, \end{aligned} \tag{14}$$

$$O(\varepsilon^2):\quad L_u(u_2, v_2) = -R_2, \quad L_v(u_2, v_2) = R_2, \tag{15}$$

$$O(\varepsilon^3):\quad L_u(u_3, v_3) = \frac{\partial u_1}{\partial T} - R_3, \quad L_v(u_3, v_3) = \eta^2 \frac{\partial v_1}{\partial T} + R_3, \tag{16}$$

where

$$R_2 = \frac{B_{cr}}{Q} u_1^2 + 2Q u_1 v_1, \quad R_3 = 2\frac{B_{cr}}{Q} u_1 u_2 + 2Q(u_1 v_2 + u_2 v_1) + u_1^2 v_1 + \mu u_1.$$

As before, the problems are considered in the entire plane $-\infty < x, y < \infty$.

We wish to describe the appearance of both hexagons and stripes as well as their interactions. Therefore, at $O(\varepsilon)$ we seek solutions of the linearized system in the form

$$\begin{pmatrix} u_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} E, \tag{17}$$

where

$$E = L_1 e_1 + L_2 e_2 + L_3 e_3 + c.c.,$$

$$e_1 = \exp(ik_{cr} x), \quad e_{2,3} = \exp\left[ik_{cr}\left(-\frac{x}{2} \pm \frac{\sqrt{3}}{2} y\right)\right], \tag{18}$$

and *c.c.* denotes complex conjugate terms. Here, the amplitudes L_1, L_2, L_3 are functions of the slow time T.

Next, we turn to the $O(\varepsilon^2)$ problem. The right-hand side R_2 can be written in the form

$$R_2 = PE^2, \quad P \equiv \frac{B_{cr}}{Q}a^2 + 2Qab,$$

and can be represented as

$$R_2 = (E_1 + E_2 + 2E_3 + 2E_4)P,$$

where

$$E_1 = L_1^2 e_1^2 + L_2^2 e_2^2 + L_3^2 e_3^2 + c.c.,$$

$$E_2 = 2(|L_1|^2 + |L_2|^2 + |L_3|^2),$$

$$E_3 = L_1 L_2^* e_1 e_2^* + L_1 L_3^* e_1 e_3^* + L_2 L_3^* e_2 e_3^* + c.c.,$$

$$E_4 = L_1 L_2 e_3^* + L_1 L_3 e_2^* + L_2 L_3 e_1^* + c.c.,$$

and the asterisk denotes the complex conjugate. We remark that the terms proportional to E_4 are secular terms that appear in the $O(\varepsilon^2)$ problem due to the resonant interaction of the modes (18). As in [54] these secular terms are considered to be small, so that they contribute to the solvability condition at $O(\varepsilon^3)$.

Then the solution of the $O(\varepsilon^2)$ problem is given by

$$\begin{pmatrix} u_2 \\ v_2 \end{pmatrix} = \left[E_1 \begin{pmatrix} u_{21} \\ v_{21} \end{pmatrix} + E_2 \begin{pmatrix} u_{22} \\ v_{22} \end{pmatrix} + 2E_3 \begin{pmatrix} u_{23} \\ v_{23} \end{pmatrix} \right] P,$$

where the coefficients u_{2j}, v_{2j} are

$$u_{21} = \frac{2^\beta k_{cr}^\beta}{(1 + 2^\alpha k_{cr}^\alpha)(Q^2 + 2^\beta k_{cr}^\beta) - 2^\beta k_{cr}^\beta B_{cr}},$$

$$v_{21} = \frac{-1 - 2^\alpha k_{cr}^\alpha}{(1 + 2^\alpha k_{cr}^\alpha)(Q^2 + 2^\beta k_{cr}^\beta) - 2^\beta k_{cr}^\beta B_{cr}},$$

$$u_{22} = 0,$$

$$v_{22} = -\frac{1}{Q^2},$$

$$u_{23} = \frac{3^{\beta/2} k_{cr}^\beta}{(1 + 3^{\alpha/2} k_{cr}^\alpha)(Q^2 + 3^{\beta/2} k_{cr}^\beta) - 3^{\beta/2} k_{cr}^\beta B_{cr}},$$

$$v_{23} = \frac{-1 - 3^{\alpha/2} k_{cr}^\alpha}{(1 + 3^{\alpha/2} k_{cr}^\alpha)(Q^2 + 3^{\beta/2} k_{cr}^\beta) - 3^{\beta/2} k_{cr}^\beta B_{cr}}.$$

We now turn to the $O(\varepsilon^3)$ problem. Substituting the solutions u_1, u_2, v_1, v_2 into R_3 yields

$$R_3 = 2PK_1EE_1 + 2PK_2EE_2 + 4PK_3EE_3 + a^2bE^3 + \mu aE,$$

where

$$K_1 = \frac{B_{cr}}{Q}au_{21} + Qav_{21} + Qbu_{21},$$

$$K_2 = \frac{B_{cr}}{Q}au_{22} + Qav_{22} + Qbu_{22},$$

$$K_3 = \frac{B_{cr}}{Q}au_{23} + Qav_{23} + Qbu_{23}.$$

The secular terms in the above products EE_1, EE_2, and so on are given as follows:

$$\text{in } EE_1: \qquad L_1|L_1|^2e_1 + L_2|L_2|^2e_2 + L_3|L_3|^2e_3 + c.c. \equiv E_0,$$

$$\text{in } EE_2: \qquad 2EF, \quad F = |L_1|^2 + |L_2|^2 + |L_3|^2,$$

$$\text{in } EE_3: \qquad EF - E_0,$$

$$\text{in } E^3: \qquad 6EF - 3E_0.$$

Thus, the secular terms in R_3 are

$$\mu aE + a\frac{\partial E}{\partial T} + E_0(2PK_1 + 4PK_2 + 3a^2b) + (EF - E_0)(4PK_2 + 4PK_3 + 6a^2b).$$

Elimination of secular terms in the $O(\varepsilon^3)$ problem results in the following system of equations for the leading order amplitudes L_1, L_2, L_3:

$$C_0\frac{dL_1}{dT} = \mu C_1L_1 + C_2L_2^*L_3^* + C_3L_1|L_1|^2 + C_4L_1(|L_2|^2 + |L_3|^2), \tag{19a}$$

$$C_0\frac{dL_2}{dT} = \mu C_1L_2 + C_2L_1^*L_3^* + C_3L_2|L_2|^2 + C_4L_2(|L_1|^2 + |L_3|^2), \tag{19b}$$

$$C_0\frac{dL_3}{dT} = \mu C_1L_3 + C_2L_1^*L_2^* + C_3L_3|L_3|^2 + C_4L_3(|L_1|^2 + |L_2|^2). \tag{19c}$$

The coefficients C_k, $k = 0, 1, 2, 3, 4$, are given by

$$C_0 = \frac{a(1+x) + \eta^2bsx}{1 + (1-s)x} = \frac{(1+x)(x^{1/s} - \eta^2sx)}{1 + (1-s)x},$$

$$C_1 = a = x^{1/s} > 0,$$

$$C_2 = 2P = \frac{2(1+x)[1+(1-2s)x]x^{2/s}}{\big[s[1+(1-s)x]x^{1+1/s}\big]^{1/2}}, \tag{20}$$

$$C_3 = 2PK_1 + 4PK_2 + 3a^2b = \frac{(1+x)x^{2/s}}{s\,x^2}\,\frac{C_{3\,n}}{C_{3\,d}},$$

$$C_4 = 4PK_2 + 4PK_3 + 6a^2b = \frac{(1+x)x^{2/s}}{s\,x^2}\,\frac{C_{4\,n}}{C_{4\,d}},$$

where

$$C_{3\,n} = \big[(-6+11s-4s^2)2^{\beta} + (6s-9s^2)2^{\alpha} + (4-9s+5s^2)2^{\alpha+\beta}\big]\,x^3$$

$$+\big[(-14+13s+s^2)2^{\beta} + 6\,s\,2^{\alpha} + (8-9s)\,2^{\alpha+\beta} + 6s - 9s^2\big]\,x^2$$

$$+\big[(-10+2s)2^{\beta} + 4\cdot 2^{\alpha+\beta} + 6s\big]\,x - 2^{1+\beta},$$

$$C_{3\,d} = \big[2^{\beta} + (s-1)2^{\alpha+\beta} - s\,2^{\alpha}\big]\,x + (1+s)2^{\beta} - 2^{\alpha+\beta} - s,$$

$$C_{4\,n} = \Big[(-8+14s-8s^2)\,3^{\beta/2} + (4-6s+2s^2)\,3^{(\alpha+\beta)/2} + (8s-10s^2)\,3^{\alpha/2}\Big]\,x^3$$

$$+\Big[(-20+22s-6s^2)\cdot 3^{\beta/2} + (8-6s)3^{(\alpha+\beta)/2} + 8\,s\,3^{\alpha/2} + 8s - 10s^2\Big]\,x^2$$

$$+\Big[(8s-16)3^{\beta/2} + 4\cdot 3^{(\alpha+\beta)/2} + 8\,s\Big]\,x - 4\cdot 3^{\beta/2},$$

$$C_{4\,d} = \Big[(-1+s)3^{(\alpha+\beta)/2} + 3^{\beta/2} - s\,3^{\alpha/2}\Big]\,x - 3^{(\alpha+\beta)/2} + (s+1)3^{\beta/2} - s.$$

We note that both C_0 and C_1 are positive. Indeed, upon simple manipulations that use (10), we obtain

$$C_0 = \frac{(1+x)sx}{1+(1-s)x}\left[\frac{x^{1/s}}{sx} - \eta^2\right]$$

$$> \frac{(1+x)sx}{1+(1-s)x}\left[\frac{x^{1/s}}{sx} - \frac{sx^{1/s}}{x+1+s}\right] = \frac{x^{1/s}(1+s)(1+x)}{1+x+s} > 0.$$

The fact that C_0 and C_1 are positive will be used in the next section.

5. Analysis of the amplitude equations: Stripes and hexagons. We consider steady states of the system (19); specifically, we are interested in the steady states that describe hexagons and stripes in the original system. We briefly list below well-known general results concerning these patterns [54] and then relate these results to the problem at hand. Hexagonal patterns correspond to

$$L_1 = L_2 = L_3 = L_h,$$

where L_h is a solution of the quadratic equation

$$(C_3 + 2C_4)L_h^2 + C_2 L_h + \mu C_1 = 0. \tag{21}$$

Stripes parallel to the y-axis correspond to

$$L_1 = L_s, \quad L_2 = L_3 = 0, \quad L_s = \sqrt{-\frac{\mu C_1}{C_3}}.$$

We first discuss the stripes. The linear stability analysis of the system (19) in the case of stripes results in the following values for the growth rate σ of perturbations:

$$C_0\sigma_1 = -2\mu C_1, \quad \sigma_2 = 0, \quad C_0\sigma_{3,4} = C_2 L_s - (C_3 - C_4)L_s^2,$$

$$C_0\sigma_{5,6} = -C_2 L_s - (C_3 - C_4)L_s^2.$$

Since the coefficients C_0 and C_1 are always positive, we conclude that supercritical stripes exist if $C_3 < 0$. They are stable if $C_3 - C_4 > 0$ and the amplitude L_s is sufficiently large, $L_s > |C_2|/(C_3 - C_4)$.

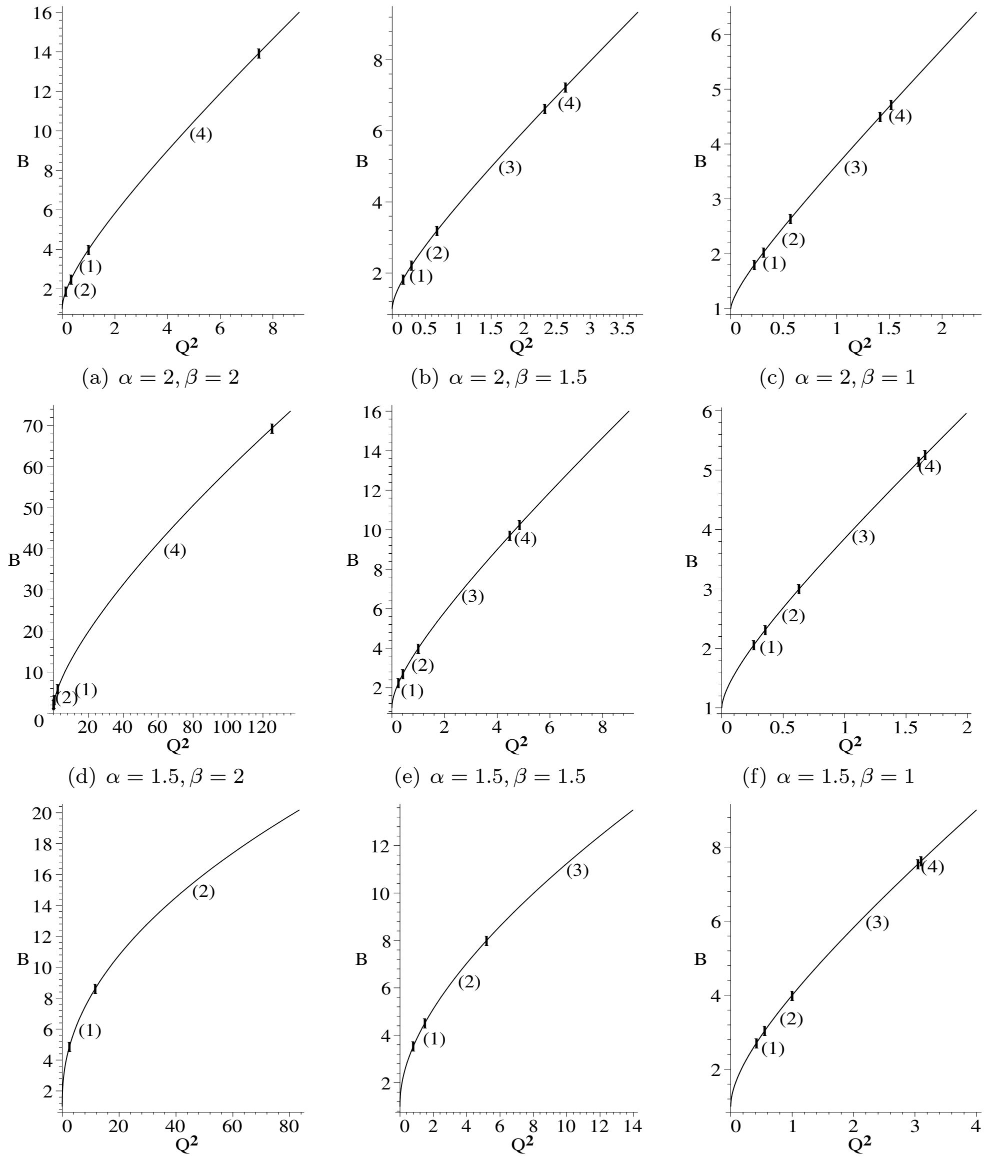
The linear stability analysis of the system (19) in the case of hexagons results in the following values for the growth rate σ of perturbations:

$$C_0\sigma_1 = C_2 L_h + 2(C_3 + 2C_4)L_h^2, \quad C_0\sigma_{2,3} = 2[-C_2 L_h + (C_3 - C_4)L_h^2],$$

$$C_0\sigma_4 = -3C_2 L_h, \quad \sigma_{5,6} = 0.$$

Again, keeping in mind that $C_0 > 0$, $C_1 > 0$, we conclude that a necessary condition for stable hexagons to exist is $C_3 + 2C_4 < 0$. This condition is not sufficient. If, in addition to this condition, $C_3 - C_4 < 0$, then the entire increasing branch of hexagons is stable. If the necessary condition is satisfied but $C_3 - C_4 > 0$, hexagons can still be stable. Specifically, they are stable if $C_3 + C_4 < 0$, but in this case only part of the increasing hexagon branch is stable. The hexagons in this case will coexist with stripes if $C_3 < 0$. Finally, we remark that in all of the above cases the hexagons may be either positive, which corresponds to $L_h > 0$ and is the case when $C_2 > 0$, or negative, which corresponds to $L_h < 0$ (for $C_2 < 0$).

In order to illustrate the effect of superdiffusion on Turing pattern formation, we consider three values of the superdiffusion exponents α and β, namely, 1, 3/2, and 2, and consider all possible combinations. In each case we plot the Turing instability boundary in the (Q^2, B) parameter plane and mark different parts of the boundary according to the different scenarios predicted by the weakly nonlinear analysis. The results are presented in Figure 3. Different parts of the stability boundaries are marked by different numbers in the figure. Crossing the boundary corresponding to different numbers as B increases leads to the appearance of different spatial patterns.



(a) $\alpha = 2, \beta = 2$

(b) $\alpha = 2, \beta = 1.5$

(c) $\alpha = 2, \beta = 1$

(d) $\alpha = 1.5, \beta = 2$

(e) $\alpha = 1.5, \beta = 1.5$

(f) $\alpha = 1.5, \beta = 1$

(g) $\alpha = 1, \beta = 2$

(h) $\alpha = 1, \beta = 1.5$

(i) $\alpha = 1, \beta = 1$

FIG. 3. *Stability boundaries in the* (Q^2, B) *plane; see text.*

Positive hexagons bifurcate when the part of the boundary marked by (1) is crossed; (2) indicates coexistence of positive hexagons and stripes, (3) shows the coexistence of negative hexagons and stripes; and finally, (4) corresponds to negative hexagons. There are parts of the boundaries that are not marked, which means that neither hexagons nor stripes are observed as these boundaries are crossed. This may occur, e.g., when both hexagons and stripes are unstable.

As Figure 3 illustrates, most of the differences between the normal diffusion and superdiffusion cases are quantitative rather than qualitative. This can be explained by the fact that in the case of a shortwave instability, the local structure of the dispersion

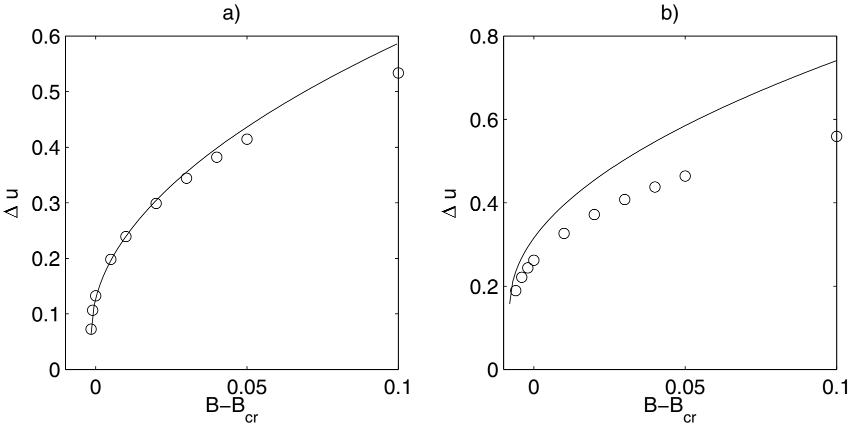


FIG. 4. *Bifurcation diagrams for hexagonal patterns. In both figures* $\alpha = \beta = 1$*;* $x = 0.9$ *in* (a), $x = 0.8$ *in* (b). *See text.*

curve near the instability threshold (parabola) remains the same as in the case of the normal diffusion. Therefore, the effect of the anomalous diffusion reduces to changing characteristic diffusion times, which leads to the renormalization of the coefficients in the amplitude equations. All types of bifurcating regimes are present in the three cases when $\beta = 1$, while larger β, e.g., $\beta = 2$, demonstrates the presence of "preferred" regimes. Another observation is that for $\alpha > \beta$ the bifurcations occur in the range of Q significantly smaller than in the case $\alpha < \beta$.

We now return to the discussion of the quadratic terms in the amplitude equations. The way the terms were treated, though commonly used [54], can by no means be claimed systematic. There is, however, a systematic way to deal with the quadratic terms. Indeed, if we consider the distinguished limit in which

$$x - \frac{1}{2s-1} = O(\varepsilon), \tag{22}$$

then $C_2 = O(\varepsilon)$ (see the expression (20) for C_2). In this case the quadratic terms become $O(\varepsilon^3)$, so that they are shifted to the $O(\varepsilon^3)$ problem, and the amplitude equations (19) arise in a systematic way due to solvability conditions applied to the $O(\varepsilon^3)$ problem. Not only is this approach more systematic, but also the results obtained under the condition (22) are expected to be more accurate. This is illustrated by comparing numerical and analytical bifurcation curves in Figure 4. Shown are numerically (circles) and analytically (solid lines) computed amplitude of u, denoted by Δu, as a function of supercriticality $B - B_{cr}$. The numerical solution was obtained for the original nondimensional problem (6) with parameter values given by (9) (details of our numerical approach are in the next section), with

$$\Delta u = \max_{(x,y)} u - \min_{(x,y)} u.$$

To obtain the analytical amplitude Δu, we used the leading order analytical solution, so that

$$\Delta u = L_h a \left[\max_{(x,y)} (e_1 + e_2 + e_3 + c.c.) - \min_{(x,y)} (e_1 + e_2 + e_3 + c.c.) \right],$$

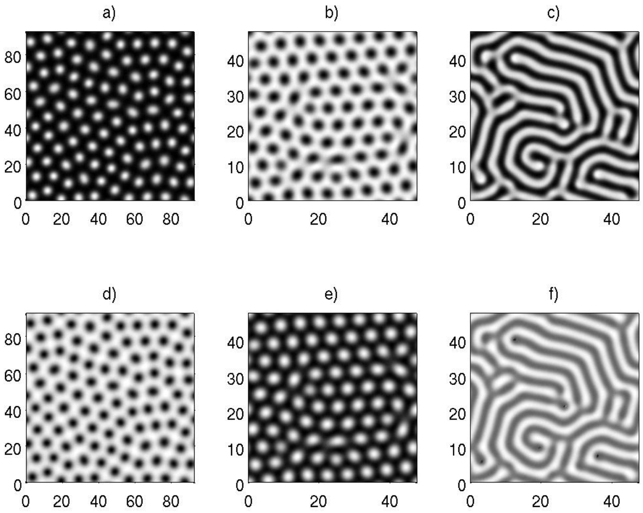


FIG. 5. *Numerical solutions of the system* (6) *in the case of superdiffusion* ($\alpha = \beta = 1.5$, $\eta = 0.2$) *showing positive hexagons* ((a) *shows* u, (d) *shows* v, $x = 0.55$, $\epsilon = 0.1$), *negative hexagons* ((b) *shows* u, (e) *shows* v, $x = 1.5$, $\epsilon = 0.1$), *and stripes* ((c) *shows* u, (f) *shows* v, $x = 1.5$, $\epsilon = 1.0$).

with L_h computed from the quadratic equation (21) as

$$L_h = \frac{1}{2(C_3 + 2C_4)} \left(-C_2 - \sqrt{C_2^2 - 4(C_3 + 2C_4)(B - B_{cr})} \right).$$

Since $\alpha = \beta = 1$ is used for the calculations shown in the figure, (22) reduces to $x - 1 = O(\varepsilon)$. For $x = 0.9$ the figure illustrates almost perfect agreement between the analytical and numerical results. For $x = 0.8$ the accuracy is acceptable. For x farther away from 1, the difference between the numerical and analytical calculations increases. However, there still is a qualitative agreement between the two, and this is why the amplitude equations (19) may be useful even if (22) is not satisfied.

6. Numerical computation of patterns. We have performed numerical computations of the system (6) by means of a pseudospectral method, with time integration in Fourier space, using a Crank–Nicolson scheme for the linear operator and an Adams–Bashforth scheme for the nonlinear operator. Periodic boundary conditions and small-amplitude random initial data have been used.

Our numerical computations confirmed the results of the weakly nonlinear analysis summarized in Figure 3. Figure 5 shows examples of positive hexagons (Figure 5(a),(d)), negative hexagons (Figure 5(b),(e)), and stripes (Figure 5(c),(f)) for the anomalous diffusion case. Note that in the hexagonal patterns several penta-hepta defects are clearly visible, while in the stripe patterns typical defects such as dislocations and disclinations are present. The defects can be described analytically as solutions of equations which generalize our equations (19) to the case when spatial modulations are present, i.e., Ginzburg–Landau equations. These defects gradually disappear as time goes on, so that the resulting final patterns will be perfect hexagons or stripes.

We also performed numerical computations of a regime where neither hexagons nor stripes are stable. The most interesting is a regime of self-replicating spots, first

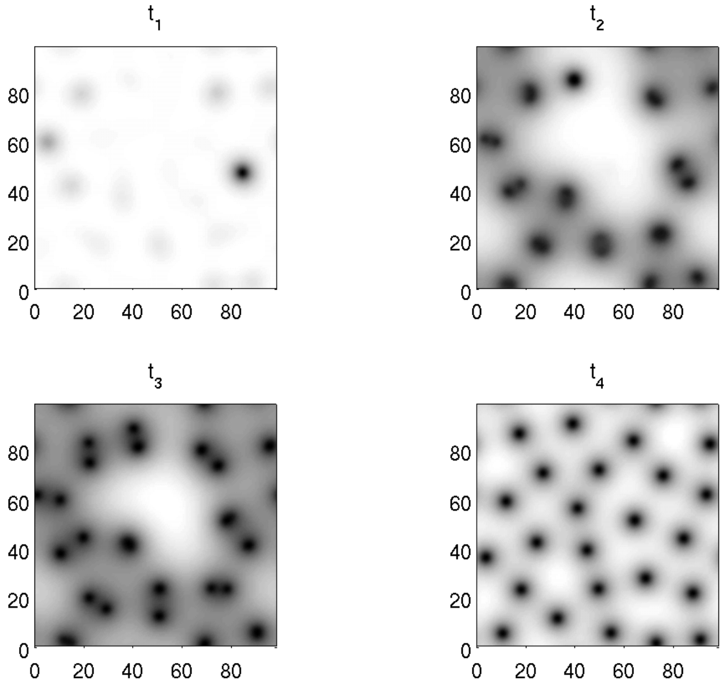


Fig. 6. *Numerical solutions of the system* (6) *in the case of normal diffusion* ($\alpha = \beta = 2.0$, $\eta = 0.2$, $x = 0.1$, $\epsilon = 0.1$) *showing the formation of self-replicating spots for the v-component. Different figures correspond to different moments of time.*

discovered in [40] for the case of normal diffusion. Figure 6 shows this regime for the case of normal diffusion. One can see that the spots appear from small perturbations of the homogeneous state and then start self-replicating until the average distance between the spots reaches a particular value. After that the spots move slowly, repelling each other, until they form a hexagonal pattern corresponding to a "Wigner crystal" [40] (not shown here).

In the case of anomalous diffusion, this regime exhibits slightly different dynamics. Namely, a single spot that appears from a small fluctuation first forms a *ring* which in turn becomes unstable and disintegrates into spots that continue to self-replicate. The ring radius at which it becomes unstable and the instability mode depend on the Lévy flight exponents α and β. Figure 7 shows the formation of a ring and its subsequent disintegration into localized spots in the case of Lévy flights with $\alpha = \beta = 1.5$. One can see that the instability occurs when the ring radius is still quite small, and the instability mode corresponds to the superposition of two modes with azimuthal numbers $m = 2$ and $m = 4$. As a result each ring produces four spots.

We have observed that the number of spots resulting from the instability of a ring depends on the anomalous diffusion exponents. The farther the anomalous diffusion exponents are from the normal diffusion case, the more spots are produced from each ring. Figure 8 shows the results of the corresponding numerical simulations for different values of α and β. One can see that for $\alpha = \beta = 1.8$ there are two rings that disintegrate into two and three spots. For $\alpha = \beta = 1.35$ the rings develop into four and five spots. Disintegration into five and seven spots can be seen for $\alpha = \beta = 1.3$.

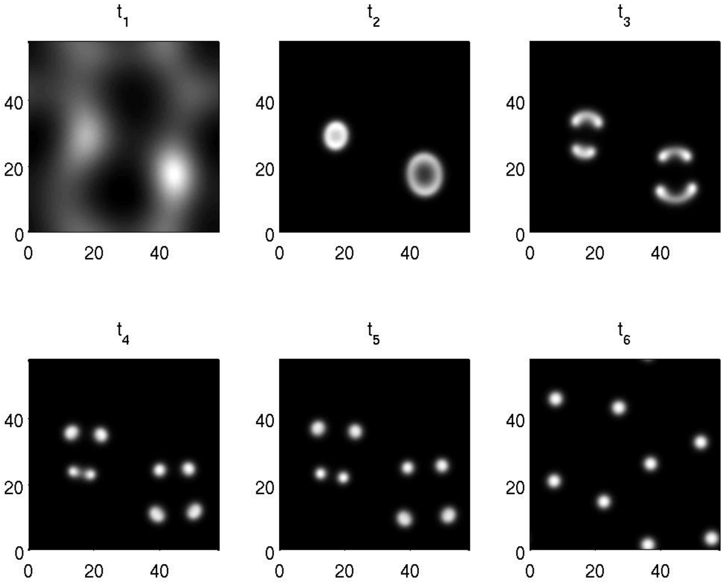


FIG. 7. *Numerical solutions of the system* (6) *in the case of superdiffusion (*$\alpha = \beta = 1.5$*,* $\eta = 0.2$*,* $x = 0.1$*,* $\epsilon = 0.1$*) showing the formation of rings that decay into localized spots. The component* u *is shown. Different figures correspond to different moments of time.*

For $\alpha = \beta = 1.2$ the ring instability results in the formation of 13 spots. This behavior may be attributed to the nonlocal character of Lévy flights: with the decrease of the anomalous exponents, different parts of the ring are "communicating" with each other more easily, which causes the ring to disintegrate into a larger number of spots.

Figure 9 shows the same type of instability in the case $\alpha = \beta = 1.0$. One can see that the ring radius at which instability occurs is much larger, and the instability mode number is much larger as well. Note too that the spots that result from the ring instability self-replicate at a later time. Finally, we note that the ring formation can be observed not only for equal Lévy exponents, as shown in Figures 7–9, but also when α and β are different but sufficiently smaller than 2.0.

7. Conclusion. The Turing mechanism of pattern formation, which is due to diffusion-induced instability of the homogeneous steady state, requires that the ratio of the diffusion coefficient of the inhibitor to that of the activator be sufficiently large in "normal" reaction-diffusion systems. One might expect, based on this result for the Brusselator problem with normal diffusion, that if the fractional derivative order β for the inhibitor Y is larger than α, the fractional derivative order for the activator (i.e., the inhibitor diffuses slower than the activator, $s < 1$), then no Turing instability can be observed. In fact, the result is quite the opposite. Not only can Turing instability occur in the case $s < 1$, which is unexpected because diffusion of the inhibitor in this case is slower than diffusion of the initiator, but it may also occur for any values of the diffusion coefficients provided that the "supply" parameter A is sufficiently large. These are the most striking qualitative differences between normal and anomalous diffusion.

The remaining results derived from the weakly nonlinear analysis indicate that

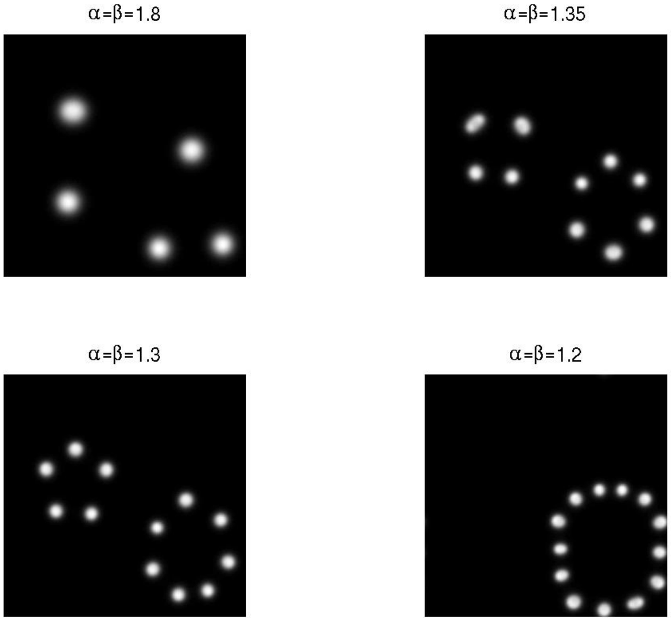


FIG. 8. *Numerical solutions of the system* (6) *in the case of superdiffusion* ($\eta = 0.2$, $x = 0.1$, $\epsilon = 0.1$) *for different values of* α *and* β *showing the result of a ring instability leading to different numbers of localized spots. The component* u *is shown.*

most of the differences between normal diffusion and superdiffusion are quantitative rather than qualitative. That is, the value of the transition point changes as normal diffusion is replaced by anomalous diffusion. This can be explained by the fact that in the case of a short-wave instability the local structure of the dispersion curve near the instability threshold is the same in both cases of normal and anomalous diffusion. The anomalous diffusion in this case just leads to the change in coefficients characterizing the parabolic character of the dispersion curve due to the change of the characteristic diffusion times. A more profound, qualitative effect of the anomalous diffusion can be expected in the case of a long-wave instability when the anomalous diffusion qualitatively changes the relation between characteristic slow spatial and time scales [34]. The analysis of this case is, however, beyond the scope of the present paper. All types of bifurcating regimes, i.e., positive and negative hexagons and regions of coexistence of these hexagons and stripes, are present when $\beta = 1$, while larger β, e.g., $\beta = 2$, demonstrates the presence of "preferred" regimes. Another observation is that for $\alpha > \beta$ the bifurcations occur in the range of A significantly smaller than in the case $\alpha < \beta$.

Our numerical computations confirmed the results of the weakly nonlinear analysis. In addition, a regime of localized, self-replicating spots has been found. We have observed that, unlike the case of normal diffusion, in the case of Lévy flights, the spots result from the instability of *localized rings.* We have exhibited the initial breakup of the rings into different numbers of spots depending on the anomalous diffusion exponents. The spots resulting from a ring instability may later self-replicate. We have observed that the number of spots resulting from the ring instability increases with the decrease of the Lévy flight exponent. This can be attributed to the increase of the nonlocal character of the superdiffusion in this case, i.e., the increase of the

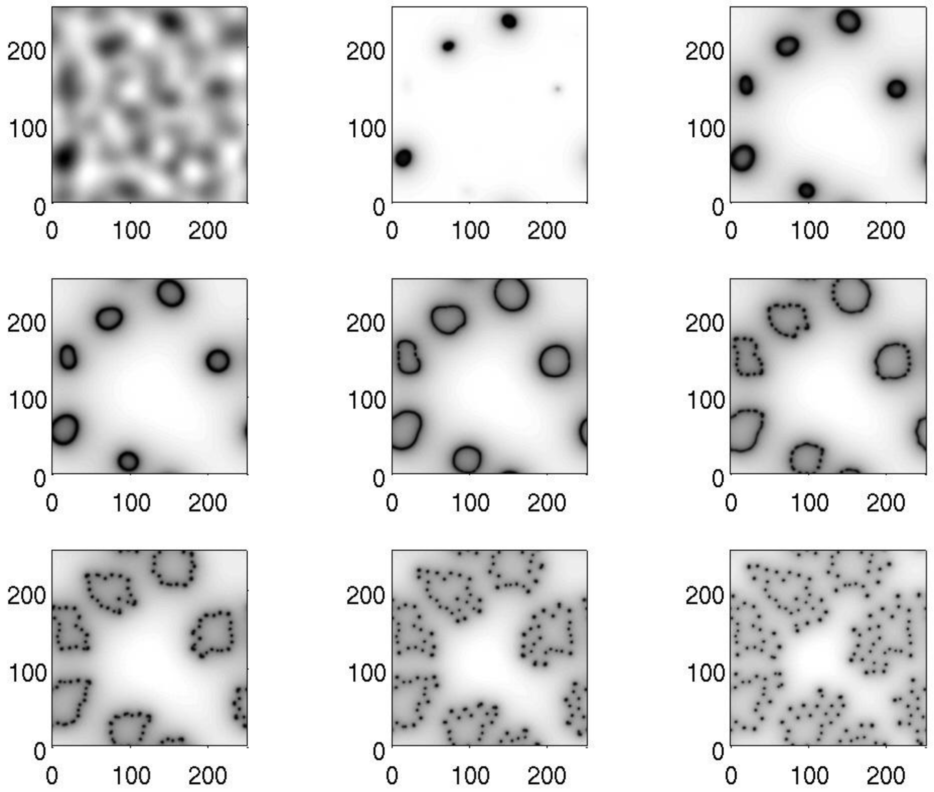


Fig. 9. *Numerical solutions of the system* (6) *in the case of superdiffusion (*$\alpha = \beta = 1.0$*,* $\eta = 0.2$*,* $x = 0.1$*,* $\epsilon = 0.1$*) showing the formation of rings that decay into localized spots. Six rings are clearly seen. The component* v *is shown. Different figures correspond to different moments of time increasing from left to right and from top to bottom.*

characteristic "communication" distance between the structures.

Finally, we note that it would be interesting to investigate the means to experimentally control the anomalous diffusion exponents in superdiffusive reaction-diffusion systems. In a system where superdiffusion is caused by, say, turbulent mixing, a possible way to control the anomalous exponents would be to vary the mixing intensity [63]. For a reaction-diffusion system in a porous medium, the control can be achieved by changing the velocity of the gas flowing through the medium. For, say, a catalytic system governed by surface diffusion and reactions, when the reactants can perform long jumps along the catalytic surface, the parameters of the surface diffusion could possibly be controlled by irradiation of the catalyst. Note that feedback control of catalytic patterns by light has been attracting a great deal of attention [32]. Another possibility would be to affect the characteristics of the anomalous surface diffusion by changing the parameters of a turbulent gas flow in the adjacent gas phase. The investigation of ways to control reaction-diffusion systems with anomalous diffusion, as well as possible feedback control mechanisms to stabilize certain desired patterns, will be considered in future studies.

REFERENCES

[1] F. Amblard, A. C. Maggs, B. Yurke, A. N. Pargellis, and S. Leibler, *Subdiffusion and anomalous local viscoelasticity in actin networks*, Phys. Rev. Lett., 77 (1996), pp. 4470–4473.

[2] R. Angelico, A. Ceglie, U. Olsson, G. Palazzo, and L. Ambrosone, *Anomalous surfactant diffusion in a living polymer system*, Phys. Rev. E, 74 (2006), 031403.

[3] J. P. Bouchaud and A. Georges, *Anomalous diffusion in disordered media—statistical mechanisms, models and physical applications*, Phys. Rep., 195 (1990), pp. 127–293.

[4] N. F. Britton, *Reaction-Diffusion Equations and Their Applications to Biology*, Academic Press, London, 1986.

[5] D. Brockmann and L. Hufnagel, *Front propagation in reaction-superdiffusion dynamics: Taming Levy flights with fluctuations*, Phys. Rev. Lett., 98 (2007), 178301.

[6] B. A. Carreras, V. E. Lynch, and G. M. Zaslavsky, *Anomalous diffusion and exit time distribution of particle tracers in plasma turbulence model*, Phys. Plasmas, 8 (2001), pp. 5096–5103.

[7] D. Del-Castillo-Negrete, B. A. Carreras, and V. E. Lynch, *Front dynamics in reaction-diffusion systems with Levy flights: A fractional diffusion approach*, Phys. Rev. Lett., 91 (2003), 018302.

[8] G. Drazer and D. H. Zanette, *Experimental evidence of power-law trapping time distributions in porous media*, Phys. Rev. E, 60 (1999), pp. 5858–5864.

[9] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer, Berlin, New York, 1979.

[10] V. V. Gafiychuk and B. Y. Datsko, *Pattern formation in a fractional reaction-diffusion system*, Phys. A, 365 (2006), pp. 300–306.

[11] H. Haken, *Synergetics: An Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*, Springer, New York, 1983.

[12] A. E. Hansen, D. Marteau, and P. Tabeling, *Two-dimensional turbulence and dispersion in a freely decaying system*, Phys. Rev. E, 58 (1998), pp. 7261–7271.

[13] J. W. Haus and K. W. Kehr, *Diffusion in regular and disordered lattices*, Phys. Rep., 150 (1987), pp. 263–406.

[14] B. I. Henry, T. A. M. Langlands, and S. L. Wearne, *Turing pattern formation in fractional activator-inhibitor systems*, Phys. Rev. E, 72 (2005), 026101.

[15] B. I. Henry and S. L. Wearne, *Existence of Turing instabilities in a two-species fractional reaction-diffusion system*, SIAM J. Appl. Math., 62 (2002), pp. 870–887.

[16] R. Hilfer, ed., *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.

[17] G. Karolyi and T. Tel, *Effective dimensions and chemical reactions in fluid flows*, Phys. Rev. E, 76 (2007), 046315.

[18] A. Kolmogoroff, I. Petrovsky, and N. Piscounoff, *Ètude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problem biologique*, Moscow Univ. Bull. Math., 1 (1937), pp. 1–25.

[19] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer, Berlin, New York, 1984.

[20] T. A. M. Langlands, B. I. Henry, and S. L. Wearne, *Turing pattern formation with fractional diffusion and fractional reactions*, J. Phys. Condens. Matter, 19 (2007), 065115.

[21] I. Lengyel and I. R. Epstein, *Modeling of Turing structures in the chlorite-iodide-malonic acid-starch reaction system*, Science, 251 (1991), pp. 650–652.

[22] I. Lengyel and I. R. Epstein, *A chemical approach to designing Turing patterns in reaction-diffusion systems*, Proc. Nat. Acad. Sci. USA, 89 (1992), pp. 3977–3979.

[23] B. W. Li and J. Wang, *Anomalous heat conduction and anomalous diffusion in one-dimensional systems*, Phys. Rev. Lett., 91 (2003). art. 044301.

[24] Y. Luo and I. R. Epstein, *Stirring and premixing effects in the oscillatory chlorite-iodide reaction*, J. Chem. Phys., 85 (1986), pp. 5733–5740.

[25] P. Manandhar, J. Jang, G. C. Schatz, M. A. Ratner, and S. Hong, *Anomalous surface diffusion in nanoscale direct deposition processes*, Phys. Rev. Lett., 90 (2003), 115505.

[26] R. Mancinelli, D. Vergni, and A. Vulpiani, *Superfast front propagation in reactive systems with non-Gaussian diffusion*, Europhys. Lett., 60 (2002), pp. 532–538.

[27] R. Mancinelli, D. Vergni, and A. Vulpiani, *Front propagation in reactive systems with anomalous diffusion*, Phys. D, 185 (2003), pp. 175–195.

[28] M. Menzinger and P. Jankowski, *Heterogeneities and stirring effects in the Belousov–Zhabotinsky reaction*, J. Phys. Chem., 90 (1986), pp. 1217–1219.

[29] R. Metzler and J. Klafter, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.

[30] R. Metzler and J. Klafter, *The restaurant at the end of the random walk: Recent developments in the description of anomalous transport by fractional dynamics*, J. Phys. A, 37 (2004), pp. R161–R208.

[31] A. S. Mikhailov, *Foundations of Synergetics* 1: *Distributed Active Systems*, Springer, Berlin, New York, 1990.

[32] A. S. Mikhailov and K. Showalter, *Control of waves, patterns and turbulence in chemical systems*, Phys. Rep., 425 (2006), pp. 79–194.

[33] J. D. Murray, *Mathematical Biology*, Springer, Berlin, New York, 1993.

[34] Y. Nec, A. A. Nepomnyashchy, and A. A. Golovin, *Oscillatory instability in super-diffusive reaction-diffusion systems: Fractional amplitude and phase diffusion equations*, Europhys. Lett., 82 (2008), 58003.

[35] Z. Neufeld, C. Lopez, E. Hernandez-Garcia, and O. Piro, *Excitable media in open and closed chaotic flows*, Phys. Rev. E, 66 (2002), 066208.

[36] Z. Neufeld, C. Lopez, E. Hernandez-Garcia, and T. Tel, *Multifractal structure of chaotically advected chemical fields*, Phys. Rev. E, 61 (2000), pp. 3857–3866.

[37] V. Nikora, H. Habersack, T. Huber, and I. McEwan, *On bed particle diffusion in gravel bed flows under weak bed load transport*, Water Resources Research, 38 (2002), p. 1081.

[38] Z. Noszticzius, Z. Bodnar, L. Garamszegi, and M. Wittmann, *Hydrodynamic turbulence and diffusion-controlled reactions: Simulation of the effect of stirring on the oscillating Belousov-Zhabotinskii reaction with the Radicalator model*, J. Phys. Chem., 95 (1991), pp. 6575–6580.

[39] Q. Ouyang and H. L. Swinney, *Transition from a uniform state to hexagonal and striped Turing patterns*, Nature, 352 (1991), pp. 610–612.

[40] J. E. Pearson, *Complex patterns in a simple system*, Science, 261 (1993), pp. 189–192.

[41] P. Ruoff, *Excitations induced by fluctuations—an explanation of stirring effects and chaos in closed anaerobic classical Belousov-Zhabotinsky systems*, J. Phys. Chem., 97 (1993), pp. 6405–6411.

[42] J. M. Sancho, A. M. Lacasta, K. Lindenberg, I. M. Sokolov, and A. H. Romero, *Diffusion on a solid surface: Anomalous is normal*, Phys. Rev. Lett., 92 (2004), 250601.

[43] H. Scher and E. W. Montroll, *Anomalous transit-time dispersion in amorphous solids*, Phys. Rev. B, 12 (1975), pp. 2455–2477.

[44] F. G. Schmitt and L. Seuront, *Multifractal random walk in copepod behavior*, Phys. A, 301 (2001), pp. 375–396.

[45] I. M. Sokolov, J. Klafter, and A. Blumen, *Fractional kinetics*, Phys. Today, 55 (2002), pp. 48–54.

[46] I. M. Sokolov, M. G. W. Schmidt, and F. Sagues, *Reaction-subdiffusion equations*, Phys. Rev. E, 73 (2006), 031102.

[47] T. H. Solomon, E. R. Weeks, and H. Swinney, *Observation of anomalous diffusion and Lévy flights in a two-dimensional rotating flow*, Phys. Rev. Lett., 71 (1993), pp. 3975–3978.

[48] J. Toner, Y. Tu, and S. Ramaswamy, *Hydrodynamics and phases of flocks*, Ann. Physics, 318 (2005), pp. 170–244.

[49] A. M. Turing, *The chemical basis of morphogenesis*, Philos. Trans. R. Soc. London Ser. B, 237 (1952), pp. 37–72.

[50] V. K. Vanag, *Waves and patterns in reaction-diffusion systems. Belousov-Zhabotinsky reaction in water-in-oil microemulsions*, Physics–Uspekhi, 47 (2004), pp. 923–941.

[51] V. K. Vanag and D. P. Melikhov, *Asymmetrical concentration fluctuations in the autocatalytic bromate-bromide-catalyst reaction and in the oscillatory Belousov-Zhabotinsky reaction in closed reactor—stirring effects*, J. Phys. Chem., 99 (1995), pp. 17372–17379.

[52] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley, *Levy flight search patterns of wandering albatrosses*, Nature, 381 (1996), pp. 413–415.

[53] M. Wachsmuth, W. Waldeck, and J. Langowski, *Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy*, J. Molecular Biol., 298 (2000), pp. 677–689.

[54] D. Walgraef, *Spatio-Temporal Pattern Formation*, Springer, New York, 1997.

[55] M. Weiss, *Stabilizing Turing patterns with subdiffusion in systems with low particle numbers*, Phys. Rev. E, 68 (2003), 036213.

[56] M. Weiss, H. Hashimoto, and T. Nilsson, *Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy*, Biophys. J., 84 (2003), pp. 4043–4052.

[57] G. Wilk and Z. Wlodarczyk, *Do we observe Levy flights in cosmic rays?*, Nuclear Phys. B Proc. Suppl., 75 (1999), pp. 191–193.

[58] A. T. Winfree, *Spiral waves of chemical activity*, Science, 175 (1972), pp. 634–636.

[59] S. B. Yuste, L. Acedo, and K. Lindenberg, *Reaction front in an* $A + B \to C$ *reaction-subdiffusion process*, Phys. Rev. E, 69 (2004), 036126.

[60] S. B. Yuste and K. Lindenberg, *Subdiffusion-limited reactions*, Chem. Phys., 284 (2002), pp. 169–180.

[61] A. N. Zaikin and A. M. Zhabotinsky, *Concentration wave propagation in two-dimensional liquid-phase self-oscillating system*, Nature, 225 (1970), pp. 535–537.

[62] A. M. Zhabotinskii, *Periodic course of oxidation of malonic acid in solution (investigation of the kinetics of the reaction of Belousov)*, Biophysics, 9 (1964), pp. 329–335.

[63] B. Zhao and J. Wang, *Stirring-controlled bifurcations in the* $1,4$-*cyclohexanedione-bromate reaction*, J. Phys. Chem. A, 109 (2005), pp. 3647–3651.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 273–282

PROPAGATION OF LOCAL DISTURBANCES IN REACTION DIFFUSION SYSTEMS MODELING QUADRATIC AUTOCATALYSIS[*]

XINFU CHEN[†] AND YUANWEI QI[‡]

Abstract. This article studies the propagation of initial disturbance in a quadratic autocatalytic chemical reaction in one-dimensional slab geometry, where two chemical species A, called the reactant, and B, called the autocatalyst, are involved in the simple scheme $A + B \to 2B$. Experiments demonstrate that chemical systems for which quadratic or cubic catalysis forms a key step can support propagating chemical wavefronts. When the autocatalyst is introduced locally into an expanse of the reactant, which is initially at uniform concentration, the developing reaction is often observed to generate two wavefronts, which propagate outward from the initial reaction zone. We show rigorously that with such an initial setting the spatial region is divided into three regions by the two wavefronts. In the middle expanding region, the reactant is almost consumed so that $A \approx 0$, whereas in the other two regions there is basically no reaction so that $B \approx 0$. Most of the chemical reaction takes place near the wavefronts. The detailed characterization of the concentrations is given for each of the three zones.

Key words. quadratic autocatalysis, traveling wave, propagation of local disturbance, reaction-diffusion

AMS subject classifications. 34C20, 34C25, 92E20

DOI. 10.1137/07070276X

1. Introduction.

In this paper we consider an isothermal autocatalytic chemical reaction step governed by the quadratic reaction relation

$$A + B \to 2B \quad \text{with rate} \quad kab.$$

Here, $k > 0$ is the reaction rate, and a and b are the concentrations of reactant A and autocatalyst B, respectively.

Well documented in the literature, the quadratic reaction relation has appeared in several important models of real chemical reactions, e.g., the Belousor–Zhabotinskii reaction and also gas-phase radical chain branching, oxidation reactions, such as the carbon-monoxide-oxygen reaction, and hydrogen-oxygen systems [13].

Experimental observations demonstrate the existence of propagating chemical wave fronts in unstirred chemical systems for which quadratic or cubic catalysis forms a key step [15], [25]. These wavefronts, or travelling waves, arise due to the interaction of reaction and diffusion. Quite often when a quantity of autocatalyst is added locally into an expanse of reactant, which is initially at uniform concentration, the ensuing reaction is observed to generate wavefronts which propagate outward from the initial reaction zone, consuming fresh reactant ahead of the wavefront as it propagates. This is the phenomenon to be addressed in this paper.

[*] Received by the editors September 13, 2007; accepted for publication (in revised form) March 31, 2008; published electronically October 13, 2008.
http://www.siam.org/journals/siap/69-1/70276.html

[†] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu). The work of this author was supported by National Science Foundation grant DMS-0504691.

[‡] Department of Mathematics, University of Central Florida, Orlando, FL 32816 (yqi@pegasus.cc.ucf.edu). The work of this author was supported by National Science Foundation grant DMS-0504691 during his visit to the University of Pittsburgh.

We study the following system for $u = u(x,t), v = v(x,t)$:

$$(1.1) \qquad \begin{cases} u_t - Du_{xx} = -uv & \text{in } \mathbb{R} \times (0,\infty), \\ v_t - v_{xx} = uv & \text{in } \mathbb{R} \times (0,\infty), \\ u(\cdot,0) = u_0(\cdot), \quad v(\cdot,0) = v_0(\cdot) & \text{on } \mathbb{R} \times \{0\}. \end{cases}$$

It is the result of simple scaling of the standard system

$$a_t = D_A a_{xx} - kab, \qquad b_t = D_B b_{xx} + kab,$$

with $D = D_A/D_B$.

Our basic assumptions are the following:

(A1) $D \in (0,1]$;

(A2) $u_0(x) = 1$ for all $x \in \mathbb{R}$; and

(A3) v_0 is a continuous nonnegative function having compact support, $v_0(0) > 0$.

Our main result is the following.

Theorem 1.1. *Assume* (A1)–(A3) *and let* (u,v) *be the solution of* (1.1). *Set*

$$(1.2) \qquad m(t) = 2t - 3(\log[3+t] - \log 3).$$

Then for each $t > 0$ *and* $x \in [-m(t), m(t)]$, *we have* $(u,v) \approx (0,1)$ *in the following sense:*

$$(1.3) \qquad u(x,t) \le e^{-\mu[m(t)-|x|]}, \qquad \Big|1 - v(x,t)\Big| \le \frac{C}{\sqrt{1 + m(t) - |x|}}.$$

On the other hand, when $x \in (-\infty, -m(t)] \cup [m(t), \infty)$, *we have* $(u,v) \approx (1,0)$ *in the sense that*

$$(1.4) \qquad \Big|1 - u(x,t)\Big| + v(x,t) \le C\Big\{1 + |x| - m(t)\Big\} e^{m(t)-|x|}.$$

A result somewhat similar to ours is obtained by Billingham and Needham [7] using formal asymptotic and numerical computation. There, instead of a Cauchy initial problem, an initial-boundary value problem on $(0,\infty)$ is considered, with a homogenous Newmann condition at $x = 0$. The proof we give here is rigorous.

It will be interesting to see how to generalize our result to the cubic autocatalysis reaction with nonlinear reaction term uv^2. But a number of technical difficulties need to be overcome, not least of which is a result similar to that of Bramson on the traveling speed of a scalar equation with nonlinearity $u(1-u)^2$.

The organization of this paper is as follows. Section 2 contains the analysis of u behind the reaction front. In section 3 the estimate of the front location is provided. The behavior of (u,v) after the reaction has taken place is shown in section 4.

We note in passing that unlike the single equation case, of which many excellent results have been proved in the last 30 years as exemplified by the works of Aronson and Weinberger [2], Fife and McLeod [10], Sattinger [21], and Chen and Guo [8] (the survey paper of Xin [24] provides a more detailed account on recent progress), there are very limited results on the study of traveling waves and their effect on global dynamics for parabolic systems. With the recent progress of proving the existence of traveling waves in [9] and [20], we hope to spur interest in such problems since many mathematical models in biology, most of which are reaction-diffusion systems, are deeply linked to traveling wave phenomena. We also note that systems similar to ours appear in the study of thermal-diffusive flows with advection; see [4], [16], [17], [18], [19], and [23].

2. Exponential decay of a reactant behind a reaction front. Whenever an autocatalyst presents, the chemical reaction takes place very fast; as a result, the reactant is consumed quickly and therefore experiences an exponential decay (in time). The central issue here is to find the spreading speed of the autocatalyst. Mathematically, by assuming $D \in (0, 1]$ (i.e., the reactant diffuses no faster than the autocatalyst does), we are able to find a good comparison to pin down the autocatalyst's spreading speed.

2.1. A comparison.

LEMMA 2.1. *Assume that $D \in (0, 1]$ and $u_0(x) \geq 0$, $v_0(x) \geq 0$, $u_0(x) + v_0(x) \geq 1$ for every $x \in \mathbb{R}$. Then the solution of (1.1) satisfies*

$$v(x, t) \geq \sqrt{D}\, \Phi(x, t) \quad \forall (x, t) \in \mathbb{R} \times (0, \infty),$$

where Φ is the solution of the initial value problem of the Fisher KPP (Kolmogorov–Petrovskii–Piskuno) equation

$$\Phi_t - \Phi_{xx} = \Phi - \Phi^2 \text{ in } \mathbb{R} \times (0, \infty), \qquad \Phi(\cdot, 0) = v_0(\cdot) \textit{ on } \mathbb{R} \times \{0\}. \tag{2.1}$$

Proof. Denote by $K(x, t)$ the fundamental solution to the heat operator,

$$K(x, t) := (4\pi t)^{-1/2} e^{-x^2/(4t)}.$$

Then the solution of (1.1) can be decomposed as

$$u = u^0 - u^1, \qquad v = v^0 + v^1,$$

where

$$\begin{aligned} u^0(x, t) &= \int_{\mathbb{R}} K(x - y, D\,t)\; u_0(y)\, dy, \\ v^0(x, t) &= \int_{\mathbb{R}} K(x - y, t)\; v_0(y)\, dy, \\ u^1(x, t) &= \int_0^t \int_{\mathbb{R}} K(x, y, D(t - s)\,)\; f(y, s)\, dyds, \\ v^1(x, t) &= \int_0^t \int_{\mathbb{R}} K(x - y, t - s)\; f(y, s)\, dyds, \\ f(x, t) &= u(x, t)\; v(x, t). \end{aligned}$$

Here u^0 and v^0 are the concentrations of the reactant and the autocatalyst, respectively, before chemical reaction is initiated. The quantity u^1 is the amount of reactant consumed and v^1 is the amount of autocatalyst produced in the reaction.

By the maximum principle, we know that $u \geq 0$ and $v \geq 0$, and so $f := uv \geq 0$. Upon noticing that

$$K(x, Dt) := (4\pi Dt)^{-1/2} e^{-x^2/(4Dt)} \leq (4\pi Dt)^{-1/2} e^{-x^2/(4t)} = D^{-1/2} K(x, t),$$

we see that

$$u^1(x, t) \leq D^{-1/2}\; v^1(x, t) \quad \forall\, (x, t) \in \mathbb{R} \times [0, \infty).$$

This implies that

$$u = u^0 - u^1 \geq u^0 - \frac{v^1}{\sqrt{D}} = u^0 - \frac{v - v^0}{\sqrt{D}} = \left(u^0 + \frac{v^0}{\sqrt{D}} \right) - \frac{v}{\sqrt{D}}\,.$$

Note that

$$\begin{aligned}
u^0(x,t)+\frac{v^0(x,t)}{\sqrt{D}} &= \int_{\mathbb{R}} K(y,Dt)u_0(x-y)dy + \frac{1}{\sqrt{D}}\int_{\mathbb{R}} K(y,t)v_0(x-y)dy\\
&= \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}}\Big\{e^{-\eta^2}u_0(x-2\eta\sqrt{Dt}\,)+e^{-D\eta^2}v_0(x-2\eta\sqrt{Dt}\,)\Big\}d\eta\\
&\geq \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} e^{-\eta^2}\Big\{u_0(x-2\eta\sqrt{Dt}\,)+v_0(x-2\eta\sqrt{Dt}\,)\Big\}d\eta\\
&\geq \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} e^{-\eta^2}d\eta = 1.
\end{aligned}$$

Thus,

$$\left(\frac{v}{\sqrt{D}}\right)_t - \left(\frac{v}{\sqrt{D}}\right)_{xx} = u\frac{v}{\sqrt{D}} \geq \left(1-\frac{v}{\sqrt{D}}\right)\frac{v}{\sqrt{D}}.$$

A simple comparison then gives $\Phi \leq v/\sqrt{D}$. □

2.2. Bramson's result. We denote by W the minimum speed traveling wave profile of the Fisher equation

$$\begin{gathered}
2W' + W'' + W - W^2 = 0 \quad \text{on } \mathbb{R},\\
W(-\infty)=1, \quad W(0)=1/2, \quad W(\infty)=0.
\end{gathered}$$

The following result can be derived from Bramson's work [3].

LEMMA 2.2. *Assume that v_0 is a nonnegative continuous function on $\mathbb{R}$ with compact support and $v_0(0) > 0$. Let Φ be the solution of* (2.1). *Then there exist constants z_+ and z_- such that*

$$\begin{aligned}
&\lim_{t\to\infty}\ \sup_{x>0}\Big|\Phi(x,t)-W([x-z_+-m(t)])\Big| = 0,\\
&\lim_{t\to\infty}\ \sup_{x<0}\Big|\Phi(x,t)-W([m(t)+z_--x])\Big| = 0,
\end{aligned}$$

where

$$m(t) := 2t - 3[\log(3+t)-\log 3] \quad \forall\, t>0.$$

2.3. The exponential decay of *u* in the reaction zone.

THEOREM 2.3. *Assume that $D\in(0,1]$, $u_0\geq 0, v_0\geq 0, u_0+v_0\geq 1$, and $v_0(0)>0$. Let (u,v) be the solution of* (1.1). *Then there exists a positive constant k such that*

$$v > k \quad \textit{in} \;\; Q := \{(x,t) \mid t>0, |x|<m(t)\}. \tag{2.2}$$

Consequently, with $\mu = [\sqrt{1+kD}-1]/D$, there holds

$$u(x,t) \leq \bar{u}(x,t) := e^{\mu[x-m(t)]} + e^{-\mu[m(t)+x]} \quad \forall\ t\geq 0,\ x\in\mathbb{R}.$$

Proof. First, applying the comparison lemma, Lemma 2.1, and Bramson's result, Lemma 2.2, we see that $v>k$ in Q.

Since $u \le 1$, we need only consider the function u in the set Q. When $(x,t) \in Q$, we use $v \ge k$ to calculate

$$\begin{aligned}\bar{u}_t - D\bar{u}_{xx} + v\bar{u} &\ge \bar{u}_t - D\bar{u}_{xx} + k\bar{u} \\ &= \bar{u}\left\{k - D\mu^2 - 2\mu + \frac{3\mu}{3+t}\right\} \ge \bar{u}[k - D\mu^2 - 2\mu] = 0.\end{aligned}$$

Since $\bar{u} > 1 \ge u$ on the parabolic boundary of Q, the assertion of the lemma thus follows from the parabolic comparison principle. □

3. Location of the reaction front. The comparison of v with the solution of the Fisher equation shows that the reaction front is at least as far as $\pm(2t - 3\log t)$ from the origin for large t. Here we show that the reaction front is located exactly in a vicinity of $\pm(2t - 3\log t)$.

For this, we denote

$$\hat{u}(x,t) = \min\left\{1, e^{\mu[x-m(t)]} + e^{-\mu[m(t)+x]}\right\}.$$

Then $u \le \hat{u}$. Consequently,

$$v_t - v_{xx} = uv \le \hat{u}v \quad \text{in } \mathbb{R} \times (0,\infty).$$

Hence, by Green's formula,

$$0 \le v(x,t) \le \int_{\mathbb{R}} G(x,t;y,0)\, v_0(y)\, dy,$$

where for each $(x,t) \in \mathbb{R} \times (0,\infty)$, $G(x,t;\cdot,\cdot)$ is the fundamental solution of

$$\begin{aligned} G_s + G_{yy} &= \hat{u}(y,s)\, G(x,t,y,s) \quad && \forall\, y \in \mathbb{R},\ s \in [0,t), \\ G(x,t;y,t) &= \delta(x-y) && \forall\, y \in \mathbb{R}.\end{aligned}$$

Here δ is the Dirac measure. Using Bramson's technique [3, Chapters 6 and 7], one can derive that

$$G(x,y,t,0) \le \frac{C(\mu)\, e^{t-|x-y|^2/(4t)}}{\sqrt{4\pi t}} (1 - e^{-|y|\, [|x|-m(t)+1]/t}).$$

Since v_0 has compact support, by following calculations illustrated in [3] we obtain the following.

LEMMA 3.1. *There exists a positive constant C_1 such that*

$$v(\pm[m(t)+z],t) \le C_1[1+|z|]e^{-z} \quad \forall z \in \mathbb{R},\ t > 0.$$

Note that when $u_0 \equiv 1$, we have $u^0 \equiv 1$ so that

$$|u-1| = u^1 \le D^{-1/2}v^1 \le D^{-1/2}v.$$

The estimate (1.4) thus follows from the above lemma.

4. Autocatalyst generated after reaction. We know that the two reaction fronts are near $m(t)$ and $-m(t)$. In the reaction zone $[-m(t), m(t)]$, the reactant is consumed very quickly. As the autocatalyst is assumed to diffuse no slower than the reactant, it is expected that $v \approx 1$ inside the reaction zone when reaction is completed. This section is devoted to proving this expectation.

4.1. An L^∞ estimate of v.

LEMMA 4.1. *There exists a positive constant C_2 such that*

$$v(x,t) \le C_2, \qquad |u_x| \le C_2 e^{-\mu||x|-m(t)|]} \quad \forall\, x \in \mathbb{R}, t > 0.$$

Proof. Set

$$K = \max\left\{\frac{1}{4}, \frac{1}{(\mu^2+2\mu)}\right\}.$$

Let t_0 be the constant such that

$$Ke^{-m(t_0)} = \frac{1}{4}.$$

Consider the function

$$\bar{v}(x,t) = 1 - K\bar{u} = 1 - Ke^{\mu[x-m(t)]} - Ke^{-\mu[x+m(t)]} \tag{4.1}$$

in the set

$$Q(t_0) := \{(x,t) \mid t > t_0, |x| < m(t) - m(t_0)\}.$$

Since $u < \bar{u}$ in $Q(t_0)$, we have

$$\bar{v}_t - \bar{v}_{xx} - u\bar{v} \ge \bar{v}_t - \bar{v}_{xx} - \bar{u}\bar{v} \ge K\bar{u}^2 > 0.$$

Then we have $\bar{v} \ge 1/2$ on the parabolic boundary of $Q(t_0)$. Hence, by comparison,

$$v \le M\ \bar{v} \ \text{ in } \ Q(t_0), \qquad M := \sup_{\partial Q(t_0)} v \le C_1[1 + m(t_0)]e^{m(t_0)}.$$

This estimate, together with Lemma 3.1, implies that v is uniformly bounded.

Once we know the boundedness of v, we can obtain the estimate for u_x by applying the local parabolic estimate. For each $x \in \mathbb{R}$ and $t \ge 2$,

$$\|u_x\|_{L^\infty(Q_1)} \le C(D)\Big\{\|f\|_{L^\infty(Q_2)} + \min\{\|u\|_{L^\infty(Q_2)}, \|u-1\|_{L^\infty(Q_2)}\}\Big\},$$

where

$$Q_1 = (x-1, x+1) \times (\max\{t-1, 0\}, t], \qquad Q_2 := (x-2, x+2) \times (\max\{t-2, 0\}, t].$$

Here we used, for simplicity, the assumption that $u_0 \equiv 1$ is a smooth function. □

4.2. The equilibrium state after reaction. Now we show that $v \approx 1$ in $(-m(t), m(t))$ for large t. For this purpose, we consider the function

$$w = u + v - u^0 - v^0.$$

Note that $u^0 \equiv 1$; then

$$\|v^0(\cdot,t)\|_{L^\infty(\mathbb{R})} = \left\|\int_{\mathbb{R}} K(\cdot - y, t)v_0(y)dy\right\|_{L^\infty(\mathbb{R})} = O\left(\frac{1}{\sqrt{t}}\right),$$

$$|u(x,t)| \le e^{-\mu|m(t)-x|} + e^{-\mu[x+m(t)]}.$$

We see that

$$|v-1| \le |w| + |u| + |v^0|.$$

The assertion (1.3) thus follows from the following.

LEMMA 4.2. *There exists a constant $C_2 > 0$ such that*

$$|w(x,t)| \le \frac{C_2}{\sqrt{m(t)-|x|}} \quad \forall\, x \in (-m(t), m(t)),\ t>0.$$

Proof. Note that w satisfies

$$w_t - w_{xx} = (D-1)u_{xx} \quad \text{in} \ \ \mathbb{R}\times(0,\infty), \qquad w(\cdot,0)=0.$$

Hence,

$$\begin{aligned} w(x,t) &= (D-1)\int_0^t \int_{\mathbb{R}} K(x-y,t-s)u_{yy}(y,s)dyds \\ &= (D-1)\int_0^t \int_{\mathbb{R}} K_x(x-y,t-s)u_y(y,s)dyds. \end{aligned}$$

It then follows that

$$|w(x,t)| \le C(1-D)\Big\{J(x,t)+J(-x,t)\Big\},$$

where

$$\begin{aligned} J(x,t) &= \int_0^t \int_{\mathbb{R}} |K_x(x-y,t-s)|e^{-\mu|y-m(s)|}dyds \\ &= \int_0^t \int_{\mathbb{R}} |K_x(x-y-m(t-s),s)|\ e^{-\mu|y|}dyds. \end{aligned}$$

To complete the proof, it suffices to show the following:

$$J(m(t)-z,t) \le \frac{C}{\sqrt{z}} \quad \forall\, z>0.$$

Let $z>0$ and $t>0$ be arbitrary. Note that

$$\begin{aligned} J(m(t)-z,t) &= \int_0^t \int_{\mathbb{R}} |K_x(m(t)-m(t-s)-z-y,s)|\ e^{-\mu|y|}dyds, \\ K_x(x,s) &= -\frac{xe^{-x^2/(4s)}}{4\sqrt{\pi}s^{3/2}}\ . \end{aligned}$$

We divide the integral in s into the following three intervals.

(i) $s\in[z/4,2z]$. For each fixed $y\in\mathbb{R}$, we have

$$\begin{aligned} &\int_{z/4}^{2z} |K_x(m(t)-m(s)-z-y,s)|ds \\ &\le \int_{z/4}^{2z} \frac{|m(t)-m(t-s)-y-z|}{4\sqrt{\pi}[z/4]^{3/2}} e^{-|m(t)-m(t-s)-y-z|^2/(4z)}ds. \end{aligned}$$

We use the change of variable from s to η defined by

$$\eta = \frac{m(t) - m(t-s) - z - y}{z}, \qquad d\eta = \frac{m'(t-s)}{z} ds = \frac{2 - \frac{3}{3+(t-s)}}{z} ds \geq \frac{ds}{z}.$$

We find that

$$\int_{z/4}^{\min\{2z,t\}} |K_x(m(t) - m(t-s) - z - y, s)| ds \leq \frac{2}{\sqrt{\pi z}} \int_{\mathbb{R}} \eta e^{-\eta^2} d\eta = \frac{2}{\sqrt{\pi z}}.$$

It then follows that

$$\int_{z/4}^{\min\{2z,t\}} \int_{\mathbb{R}} |K_x(m(t) - m(t-s) - z - y, s)| e^{-\mu|y|} ds dy \leq \frac{2}{\sqrt{\pi z}} \int_{\mathbb{R}} e^{-\mu|y|} dy \leq \frac{4}{\mu\sqrt{\pi z}}.$$

(ii) $s > 2z$. We write

$$\int_{\mathbb{R}} |K_x(m(t) - m(t-s) - z - y, s)| e^{-\mu|y|} dy = \int_{|y|>s/6} + \int_{|y|<s/6}.$$

For the first integral,

$$\begin{aligned} \int_{|y|>s/6} &\leq e^{-\mu s/6} \int_{\mathbb{R}} |K_x(m(t) - m(t-s) - z - y, s| dy \\ &= 2e^{-\mu s/6} K(0,s) = \frac{e^{-\mu s/6}}{\sqrt{\pi s}}. \end{aligned}$$

For the second integral, we first notice that $|m(t) - m(t-s)| \geq s$ (since $1 \leq m' < 2$ on $[0,\infty)$). Hence, when $|y| < s/6$,

$$|m(t) - m(t-s) - z - y| \geq |m(t) - m(t-s)| - z - y \geq s - \frac{s}{2} - \frac{s}{6} = \frac{s}{3}.$$

Consequently,

$$\int_{|y|<s/6} |K_x| e^{-\mu|y|} dy \leq \int_{|x|>s/3} |K_x(x,s)| dx = 2K\Big(\frac{s}{3}, s\Big) = \frac{\sqrt{3}e^{-s/36}}{\sqrt{\pi s}}.$$

Thus,

$$\begin{aligned} &\int_z^t \int_{\mathbb{R}} |K_x(m(t) - m(t-s) - z - y, s)|\ e^{-\mu|y|} dy ds \\ &\qquad \leq \int_z^\infty \left(\frac{e^{-\mu s/6}}{\sqrt{\pi s}} + \frac{\sqrt{3}e^{-s/36}}{\sqrt{\pi s}} \right) ds = O(e^{-z/36}) + O(e^{-\mu z/6}). \end{aligned}$$

(iii) $0 < s < z/4$. We write

$$\int_{\mathbb{R}} |K_x(m(t) - m(t-s) - z - y, s)| e^{-\mu|y|} dy = \int_{|y|>z/4} + \int_{|y|<z/4}.$$

The first integral is easy to estimate:

$$\int_{|y|>z/4} \leq e^{-\mu z/4} \int_{\mathbb{R}} |K_x| dy = \frac{e^{-\mu z/4}}{\sqrt{\pi s}}.$$

For the second integral, we notice that $|m(t)-m(t-s)| \le 2s \le z/2$, so that when $|y| \le z/4$, we have

$$|m(t)-m(t-s)-z-y| \ge z-|m(t)-m(t-s)|-|y| \ge z-z/2-z/4 = z/4.$$

Also, $|m(t)-m(t-s)-z-y| < 2z$. Hence,

$$|K_x(m(t)-m(t-s)-z-y,s)| \le \frac{ze^{-z^2/(64s)}}{2\sqrt{\pi}s^{3/2}}.$$

It follows that

$$\int_{|y|<z/4} |K_x|e^{-\mu|y|}dy \le \frac{ze^{-z^2/(64s)}}{2\sqrt{\pi}s^{3/2}} \int_{\mathbb{R}} e^{-\mu|y|} = \frac{ze^{-z^2/(64s)}}{\mu\sqrt{\pi}s^{3/2}}.$$

Thus,

$$\begin{aligned}\int_0^{\min\{z/4,t\}} \int_{\mathbb{R}} |K_x|e^{-\mu|y|}dy &\le \int_0^{z/4} \frac{ze^{-z^2/(64s)}ds}{\mu\sqrt{\pi}s^{3/2}} + \int_0^z \frac{e^{-\mu z/4}ds}{\sqrt{\pi s}} \\ &= \int_{\sqrt{z/8}}^{\infty} \frac{4}{\mu\sqrt{\pi}}e^{-\eta^2}d\eta + \frac{\sqrt{z}e^{-\mu z/4}}{\sqrt{\pi}} = O(e^{-\mu z/8}).\end{aligned}$$

Combining all these estimate, we then obtain the assertion of the lemma. □

Proof of Theorem 1.1. The theorem follows directly from the results of Theorem 2.3 and Lemmas 4.1 and 4.2.

REFERENCES

[1] R. Aris, P. Gray, and S. K. Scott, *Modelling of cubic autocatalysis by successive biomolecular steps,* Chem. Eng. Sci., 43 (1988), pp. 207–211.

[2] D. G. Aronson and H. F. Weinberger, *Multidimensional diffusion arising in population genetics,* Adv. Math., 30 (1978), pp. 33–76.

[3] M. Bramson, *Convergence of solutions of the Kolmogorov equation to travelling waves*, Mem. Amer. Math. Soc., 44 (1983), no. 285.

[4] H. Berestycki, F. Hamel, A. Kiselev, and L. Ryzhik, *Quenching and propagation in KPP reaction-diffusion equations with a heat loss*, Arch. Ration. Mech. Anal., 178 (2005), pp. 57–80.

[5] J. Billingham and D. J. Needham, *The development of travelling wave in quadratic and cubic autocatalysis with unequal diffusion rates.* I. *Permanent from travelling waves,* Philos. Trans. R. Soc. London Ser. A, 334 (1991), pp. 1–24.

[6] J. Billingham and D. J. Needham, *The development of travelling wave in quadratic and cubic autocatalysis with unequal diffusion rates.* II. *An initial value problem with an immobilized or nearly immobilized autocatalyst,* Philos. Trans. R. Soc. London Ser. A, 336 (1991), pp. 497–539.

[7] J. Billingham and D. J. Needham, *The development of travelling waves in quadratic and cubic autocatalysis with unequal diffusion rates.* III. *Large time development in quadratic autocatalysis*, Quart. Appl. Math., 50 (1992), pp. 343–372.

[8] X. Chen and J.-S. Guo, *Existence and asymptotic stability of traveling waves of discrete quasilinear monostable equations*, J. Differential Equations, 184 (2002), pp. 549–569.

[9] X. Chen and Y. Qi, *Sharp estimates on minimum travelling wave speed of reaction diffusion systems modelling autocatalysis*, SIAM J. Math. Anal., 39 (2007), pp. 437–448.

[10] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling wave front solutions,* Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[11] S. Focant and Th. Gallay, *Existence and stability of propagating fronts for an autocatalytic reaction-diffusion system,* Phys. D, 120 (1998), pp. 346–368.

[12] R. J. Gowland and G. Stedman, *A novel moving boundary reaction involving hydroxylamine and nitric acid*, J. Chem. Soc. Chem. Comm., 10 (1983), pp. 1038–1039.

[13] P. Gray, J. F. Griffiths, and S. K. Scott, *Experiemental studies of the ignition diagram and the effect of added hydrogen*, Proc. Roy. Soc. London Ser. A, 397 (1984), pp. 21–44.

[14] P. Gray and S. K. Scott, *Chemical Oscillations and Instabilties*, Clarendon, Oxford, 1990.

[15] A. Hanna, A. Saul, and K. Showalter, *Detailed studies of propagating fronts in the iodate oxidation of arsenous acid*, J. Amer. Chem. Soc., 104 (1982), pp. 3838–3844.

[16] S. Heinze, G. Papanicolaou, and A. Stevens, *Variational principles for propagation speeds in inhomogeneous media,* SIAM J. Appl. Math., 62 (2001), pp. 129–148.

[17] A. Kiselev and L. Ryzhik, *Enhancement of the traveling front speeds in reaction-diffusion equations with advection,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 309–358.

[18] B. Khouider, A. Bourlioux, and A. Majda, *Parametrizing the burning speed enhancement by small-scale period flows.* I. *Unsteady shears, flame residence time and bending*, Combust. Theory Model., 5 (2001), pp. 295–318.

[19] J. Nolen and J. Xin, *A variational principle based study of KPP minimal front speeds in random shears*, Nonlinearity, 18 (2005), pp. 1655–1675.

[20] Y. W. Qi, *The development of travelling waves in cubic auto-catalysis with different rates of diffusion*, Phys. D, 226 (2007), pp. 129–135.

[21] D. Sattinger, *On the stability of waves of nonlinear parabolic systems,* Adv. Math., 22 (1976), pp. 312–355.

[22] A. Saul and K. Showalter, *Propagating reaction-diffusion fronts*, in Oscillations and Traveling Waves in Chemical Systems, R. J. Field and M. Burger, eds., Wiley, New York, 1984, pp. 419–439.

[23] J. Shi and X. Wang, *Hair-triggered instability of radial steady states, spread and extinction in semilinear heat equations*, J. Differential Equations, 231 (2006), pp. 235–251.

[24] J. Xin, *Front propagation in heterogeneous media*, SIAM Rev., 42 (2000), pp. 161–230.

[25] A. N. Zaikin and A. M. Zhabotinskii, *Concentration wave propagation in two-dimensional liquid-phase self-organising systems*, Nature, 225 (1970), pp. 535–537.

SIAM J. APPL. MATH.
Vol. 69, No. 1, pp. 283–304

A THERMAL ELASTIC MODEL FOR DIRECTIONAL CRYSTAL GROWTH WITH WEAK ANISOTROPY[*]

JINBIAO WU[†], C. SEAN BOHUN[‡], AND HUAXIONG HUANG[§]

Abstract. In this paper we present a semi-analytical thermal stress solution for directional growth of type III–V compounds with small lateral heat flux and weak anisotropy. Both geometric and material anisotropy are considered, and our solution can be applied to crystals grown by various growth techniques such as the Czochralski (Cz) method. The semi-analytical nature of the solution allows us to compute thermal stress in crystals with weak anisotropic effects much more efficiently, compared to a full 3D simulation. Examples are given for crystals pulled in a variety of seed orientations. Our results show that the geometric effect is the dominant one while the effect of material anisotropy on thermal stress is secondary.

Key words. crystal growth, asymptotic expansion, anisotropy, facet formation, thermal stress, Czochralski technique

AMS subject classifications. 74A10, 74E10, 74F05, 74H10, 80A22, 82D25, 82D37

DOI. 10.1137/070698439

1. Introduction.

Directional growth techniques such as the Czochralski (Cz) method are frequently used to produce high quality single crystals. By dipping a small seed crystal into a pool of molten material in the crucible and carefully controlling the heat balance inside the grower, a large crystal can be grown by pulling the crystal away from the melt in a slow and steady fashion. The pulling rod and the crucible are normally rotated in opposite directions during the growth period. For a more detailed account of the Cz and other techniques, we refer the readers to the extremely informative handbooks by Hurle [5]. Almost perfectly cylindrical crystals are grown for silicon and other semiconductor materials despite their internal structure and material anisotropy. For these crystals, the effect of material anisotropy on thermal stress has been investigated by assuming an axisymmetric cylindrical shape [3, 9, 11, 12]. On the other hand, anisotropic effects such as facets are often visible on the surface of binary compound semiconductor crystals grown by these methods [8]. The effect of a noncylindrical shape on the thermal stress is therefore of practical interest.

In a previous paper, we developed a thermal stress model for directional growth of crystals with facets [13]. For constrained growth such as that of the Cz method, a lateral growth model consistent with the lattice structure of type III–V crystals was proposed. This model is capable of predicting facet formation on the lateral surface, which qualitatively resembles experimental observations [8] of indium antimonide (InSb) crystals. Furthermore, under the assumptions of weak lateral heat flux, we have derived perturbation solutions for temperature and related thermal stress for faceted crystals by neglecting material anisotropy.

[*] Received by the editors July 26, 2007; accepted for publication (in revised form) June 10, 2008; published electronically October 13, 2008. The authors would like to acknowledge the support from NSERC (HH), MITACS (HH and JW), NSFC-10501001 (JW), and Firebird Inc. (HH and CSB).
http://www.siam.org/journals/siap/69-1/69843.html

[†] LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, China (jwu@math.pku.edu.cn).

[‡] Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON, Canada L1H 7K4 (sean.bohun@uoit.ca).

[§] Corresponding author. Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J 1P3 (hhuang@yorku.ca).

The effect of material anisotropy on thermal stress, on the other hand, could be significant for cylindrical crystals with an underlying cubic lattice structure, as shown in [11, 12]. It is, however, not clear whether the conclusions in [11, 12] hold for InSb crystals grown in a noncylindrical shape, especially those with facets forming on the lateral surface. The purpose of this paper is to investigate the combined effect of the geometric and material anisotropy on thermal stress inside the conic crystals considered in [13]. We start with the description of the mathematical model and the thermal problem in section 2. Since the growth model and temperature solution are identical to those in [13], they are presented without detailed derivation.

The main results of this paper are given in section 3, where the detailed derivation of the thermal stress with anisotropic elastic constants is presented. We show that the thermal stress can be expanded into an asymptotic series with respect to ω, a measure of the material anisotropy, and in Appendix A we prove that the series converges. As a result, a systematic approach can be devised to compute thermal stress to arbitrary order with the zeroth order solution corresponding to the case of isotropic material constants. In section 4, we present computational results for crystals pulled in a variety of seeding orientations. The results show that the effect of material anisotropy could be significant when the geometric effect is absent. The geometric effect, when it is present, usually dominates.

2. Model.

The basic assumptions of our model are that the lateral heat flux is small and the material and geometric anisotropic effects are weak, following [1, 13]. To simplify the discussion, we assume that lateral heat transfer from the crystal to the background is known. We also assume that the heat flux from the melt is fixed while the pull rate can be adjusted in order to grow a crystal with a desirable lateral profile, e.g., a conic crystal. In principle, we could incorporate the effect of the melt flow by coupling the heat transfer process in the crystal with that in the melt. However, to focus on the thermal stress in the crystal, we neglect the effect of the melt flow and assume that the axial heat flux from the melt at the crystal/melt interface does not vary in the cross-sectional (radial and circumferential) directions.

2.1. Thermal problem.

Within the crystal Ω, the temperature $T(\mathbf{x},t)$ satisfies the heat equation,

$$\rho_s c_s \frac{\partial T}{\partial t} = \nabla \cdot (\kappa_s \nabla T), \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{2.1a}$$

where ρ_s, c_s, and κ_s are the density, specific heat, and thermal conductivity of the crystal, respectively. The boundary conditions on the crystal/gas interface Γ_g, and the chuck (holding the seed), are,

$$-\kappa_s \frac{\partial T}{\partial \mathbf{n}} = h_{\mathrm{gs}}(T - T_g) + h_F(T^4 - T_b^4), \qquad \mathbf{x} \in \Gamma_g, \tag{2.1b}$$

$$\kappa_s \frac{\partial T}{\partial z} = h_{\mathrm{ch}}(T - T_{\mathrm{ch}}), \qquad z = 0, \tag{2.1c}$$

where h_{gs} and h_{ch} represent the heat transfer coefficients of the crystal/gas and crystal/chuck interfaces; h_F is the radiation heat transfer coefficient; and T_g, T_{ch}, and T_b denote the ambient gas temperature, the chuck temperature, and background temperature, respectively.

The crystal/melt interface is denoted Γ_S and is where $T = T_m$, which is the melting temperature. Explicitly we denote the melting isotherm by

$$z - S(\mathbf{x},t) = 0, \qquad \mathbf{x} \in \Gamma_S, \tag{2.1d}$$

with S denoting the crystal/melt interface. The motion of the interface of the phase transition is governed by the Stefan condition,

$$\rho_s L|\mathbf{v}_n| = \kappa_s \left.\frac{\partial T}{\partial \mathbf{n}}\right|_{z\to S^-} - q_{l,n}, \qquad |\mathbf{v}_n| = v_n = \frac{\partial S}{\partial t}\mathbf{k}\cdot\mathbf{n}, \tag{2.1e}$$

where L is the latent heat, $|\mathbf{v}_n|$ is the speed of the interface in the direction of its outward normal $\mathbf{n}$, $q_{l,n}$ is the heat flux from the melt normal to the interface, and $\partial S/\partial t$ is the speed of the interface S in the $\mathbf{k}$ direction.

2.2. Crystal shape. For the purpose of computing thermal stress, we assume that the shape of the crystal is self-similar in that the crystal radius $R(\phi, z)$ scales with its angular average $\bar{R}(z)$. Such crystals are indeed seen experimentally [8], and representing the angular dependence with a truncated Fourier series gives

$$R(\phi, z) = \bar{R}(z)\left(1 + \epsilon\sum_{k=1}^{m}\beta_k \cos(n_k\phi + \delta_k)\right) = \bar{R}(z)\left(1 + \epsilon\lambda(\phi)\right). \tag{2.2a}$$

In this expression $2\pi\bar{R}(z) = \int_{-\pi}^{\pi} R(\phi, z)\,d\phi$, m, $n_1 < n_2 < \cdots < n_m$ are positive integers ($m = 1$, $n_1 = 4$ for fourfold symmetry); the $\delta_k \in [0, 2\pi)$ are chosen to ensure that $\beta_k > 0$. Note that only those n_k corresponding to $\beta_k > 0$ are represented in the series. The parameter $\epsilon \geq 0$ is a measure of the anisotropy, and choosing the β_k so that $\sum_{k=1}^{m}\beta_k^2 = 1$ yields the expression

$$\epsilon^2 = \frac{1}{\pi\bar{R}(z)^2}\int_{-\pi}^{\pi}\left(R(\phi, z) - \bar{R}(z)\right)^2 d\phi. \tag{2.2b}$$

We will assume that $\epsilon < 1$, which will certainly[1] be the case if $\max_\phi R(\phi, z) < \frac{3}{2}\bar{R}(z)$. The β_k, δ_k, and ϵ values for the cross sections used in section 4 can be found in [13].

Of particular interest are the angular integrals

$$I_{i,j}(\epsilon) = \int_0^{2\pi}(1+\epsilon\lambda)^i\left((1+\epsilon\lambda)^2 + (\epsilon\lambda')^2\right)^{j/2} \mathrm{d}\phi \tag{2.2c}$$

$$= 2\pi + \frac{\pi}{2}\left[(i+j)(i+j-1) + j\sum_{k=1}^{m} n_k^2\beta_k^2\right]\epsilon^2 + O(\epsilon^3), \tag{2.2d}$$

where $i, j \in \mathbb{Z}$ and $\lambda' = d\lambda/d\phi$. Both the enclosed area (A) and circumference (s) of R will be utilized in what follows. For any fixed z it is an easy exercise to verify $A(z) = \bar{R}^2 I_{2,0}/2$ and $s(z) = \bar{R}I_{0,1}$.

2.3. Nondimensionalization. For simplicity, we assume that the gas temperature T_g is constant. Defining the Biot number by

$$\mathrm{Bi} = \frac{\bar{h}_{\mathrm{gs}}\tilde{R}}{\kappa_s}, \tag{2.3}$$

where $\tilde{R}$ is a characteristic radius of the crystal and $\bar{h}_{\mathrm{gs}}$ is the mean value of h_{gs}, we adopt the following scalings:

$$r = \tilde{R}\hat{r}, \quad R(\phi, z) = \tilde{R}\hat{R}(\hat{\phi}, \hat{z}), \quad \mathrm{Bi}^{1/2}\, z = \tilde{R}\hat{z}, \quad \mathrm{Bi}^{1/2}\, S(r, \phi, t) = \tilde{R}\hat{S}(\hat{r}, \hat{\phi}, \hat{t}),$$

$$\mathrm{St} = \frac{L}{c_s\Delta T}, \quad \Delta T = T_m - T_g, \quad T = T_g + \Delta T\Theta, \quad t = \frac{\mathrm{St}\,\tilde{R}^2\rho_s c_s}{\kappa_s\,\mathrm{Bi}}\hat{t},$$

[1] This is simply a consequence of the positivity of R and the mean value theorem applied to (2.2b).

with $\phi = \hat{\phi}$. Here variables with hats ($\hat{}$) are the nondimensional ones. In terms of these variables, the heat equation (2.1a) becomes (after dropping hats)

$$\frac{\mathrm{Bi}}{\mathrm{St}}\Theta_t = \frac{1}{r}(r\Theta_r)_r + \frac{1}{r^2}\Theta_{\phi\phi} + \mathrm{Bi}\,\Theta_{zz}, \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{2.4a}$$

and boundary conditions (2.1b)–(2.1d) under the nondimensional scaling become

$$-\Theta_r + \frac{1}{R^2}R_\phi\Theta_\phi + \mathrm{Bi}\,R_z\Theta_z = \mathrm{Bi}\,F(\Theta)\left(1 + \frac{R_\phi^2}{R^2} + \mathrm{Bi}\,R_z^2\right)^{1/2}, \qquad \mathbf{x} \in \Gamma_g, \tag{2.4b}$$

$$\Theta_z(0,\phi,t) = \delta\left(\Theta(0,\phi,t) - \Theta_{\mathrm{ch}}\right), \tag{2.4c}$$

$$\Theta = 1, \qquad \mathbf{x} \in \Gamma_S, \tag{2.4d}$$

where $\Theta_{\mathrm{ch}} = (T_{\mathrm{ch}} - T_g)/\Delta T$ and

$$F(\Theta) = \frac{h_F(T_g^4 - T_b^4)}{\bar{h}_{\mathrm{gs}}\Delta T} + \left(\beta(z) + \frac{4h_F}{\bar{h}_{\mathrm{gs}}}T_g^3\right)\Theta + \frac{h_F}{\bar{h}_{\mathrm{gs}}}\Delta T(6T_g^2 + 4T_g\Delta T\Theta + \Delta T^2\Theta^2)\Theta^2,$$

$\beta(z) = h_{\mathrm{gs}}/\bar{h}_{\mathrm{gs}}$, and $\delta = \mathrm{Bi}^{1/2}\,h_{\mathrm{ch}}/\bar{h}_{\mathrm{gs}}$. The crystal/melt interface advances according to the Stefan condition (2.1e), which in nondimensional coordinates becomes

$$\Theta_z - \frac{1}{\mathrm{Bi}}S_r\Theta_r - \frac{1}{\mathrm{Bi}\,r^2}S_\phi\Theta_\phi = \gamma + S_t, \qquad \gamma = \frac{q_l\tilde{R}}{\mathrm{Bi}^{1/2}\,\kappa_s\Delta T}, \tag{2.4e}$$

where γ (q_l) is the nondimensional (resp., dimensional) heat flux in the liquid across the crystal/melt interface in the axial direction.

2.4. Temperature solution. Equations (2.4a) and (2.4b) strongly suggest that the temperature Θ is independent of r and ϕ to leading order. If true, then the crystal/melt interface S is also independent of r and ϕ to leading order. These observations motivate the following approximates:

$$\begin{aligned} \Theta &\sim \Theta_0(z,t) + \mathrm{Bi}\,\Theta_1(r,\phi,z,t) + \mathrm{Bi}^2\,\Theta_2(r,\phi,z,t) + \cdots, \\ S &\sim S_0(t) + \mathrm{Bi}\,S_1(r,\phi,t) + \mathrm{Bi}^2\,S_2(r,\phi,t) + \cdots. \end{aligned} \tag{2.5}$$

We substitute them into the scaled model, expand in powers of Bi, and simplify and collect terms of the same orders.

The zeroth order problem is given by[2]

$$\frac{1}{\mathrm{St}}\Theta_{0,t} - \Theta_{0,zz} = \frac{2}{\bar{R}}\left(\bar{R}'\Theta_{0,z} - \frac{I_{0,1}}{I_{2,0}}F(\Theta_0)\right), \qquad 0 < z < S_0(t),\ t > 0, \tag{2.6a}$$

$$\Theta_{0,z}(0,t) = \delta(\Theta_0(0,t) - \Theta_{\mathrm{ch}}), \qquad t \geq 0, \tag{2.6b}$$

$$\Theta_0(S_0(t),t) = 1, \qquad t \geq 0, \tag{2.6c}$$

$$S_0'(t) = \Theta_{0,z}(S_0(t),t) - \gamma, \qquad S_0(0) = Z_0,\ t > 0. \tag{2.6d}$$

The first order solution is given by

$$\Theta_1(r,\phi,z,t) = \Theta_1^a(z,t) + r^2\Theta_1^b(z,t) + \epsilon\Theta_1^c(r,\phi,z,t) + O(\epsilon^2), \tag{2.7a}$$

[2] For weak anisotropy, $I_{0,1}/I_{2,0} = 1 + O(\epsilon^2)$.

where, keeping only those terms to $O(\epsilon)$,

$$\Theta_1^b(z,t) = \frac{1}{2\bar{R}}\left(\bar{R}'\Theta_{0,z} - F(\Theta_0)\right), \tag{2.7b}$$

$$\Theta_1^c(r,\phi,z,t) = \bar{R}F(\Theta_0)\sum_{k=1}^{m}\frac{\beta_k}{n_k}\left(\frac{r}{\bar{R}}\right)^{n_k}\cos(n_k\phi+\delta_k). \tag{2.7c}$$

These last two terms are completely determined by Θ_0 and $\bar{R}$. The first term $\Theta_1^a(z,t)$ can be found in [13] and is not repeated here since it is not relevant to the stress computation.

3. Thermal stress. We now turn our attention to thermal stress. In the following, the general case in 3D space is discussed first, followed by a more detailed discussion using the plane-strain assumption.

3.1. Thermoelasticity equations for solids with cubic anisotropy. We consider a 3D elasticity problem for a crystal with cubic symmetry as in [6]. In this case the stresses $\boldsymbol{\sigma} = (\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{yz}, \sigma_{xz}, \sigma_{xy})^{\mathrm{T}}$ and strains $\mathbf{e} = (e_{xx}, e_{yy}, e_{zz}, 2e_{yz}, 2e_{xz}, 2e_{xy})^{\mathrm{T}}$ are related through

$$\boldsymbol{\sigma} = C\mathbf{e}, \qquad C = \begin{pmatrix} C_{11} & C_{12} & C_{12} & & & \\ C_{12} & C_{11} & C_{12} & & & \\ C_{12} & C_{12} & C_{11} & & & \\ & & & C_{44} & & \\ & & & & C_{44} & \\ & & & & & C_{44} \end{pmatrix}. \tag{3.1}$$

We denote the displacement vector by $\mathbf{w}$ and the related strain by $\mathbf{e} = \mathbf{S}(\mathbf{w})$ so that the related stress tensor is given by $\boldsymbol{\sigma} = \mathcal{D}\mathbf{S}(\mathbf{w})$. For an anisotropic material, the quantity $H = 2C_{44} - C_{11} + C_{12} \neq 0$. By defining $C = C^0 - C^a$, where $C^a = H/4 \times \operatorname{diag}(2,2,2,-1,-1,-1)$, the matrix

$$C^0 = \begin{pmatrix} C^0_{11} & C^0_{12} & C^0_{12} & & & \\ C^0_{12} & C^0_{11} & C^0_{12} & & & \\ C^0_{12} & C^0_{12} & C^0_{11} & & & \\ & & & C^0_{44} & & \\ & & & & C^0_{44} & \\ & & & & & C^0_{44} \end{pmatrix} \tag{3.2a}$$

is isotropic and the quantities E and ν in terms of C^0_{ij} are given by [10]

$$E = \frac{(C^0_{11} + 2C^0_{12})(C^0_{11} - C^0_{12})}{C^0_{11} + C^0_{12}}, \qquad \nu = \frac{C^0_{12}}{C^0_{11} + C^0_{12}}. \tag{3.2b}$$

By adopting the scaling in section 2.3 for r and T in addition to (α is the thermal expansion coefficient)

$$\mathbf{w} = \tilde{R}\alpha\Delta T\hat{\mathbf{w}}, \quad \sigma_{ij} = \frac{\alpha\Delta TE}{1-\nu}\hat{\sigma}_{\hat{i}\hat{j}}, \quad e_{ij} = \alpha\Delta T\hat{e}_{\hat{i}\hat{j}},$$

we set

$$C_{ij} = \frac{E}{1-\nu}\hat{C}_{\hat{i}\hat{j}}, \qquad H = \frac{E}{1-\nu}\hat{H}$$

and obtain (after dropping hats)

$$C_{11}^0 = \frac{(1-\nu)^2}{(1+\nu)(1-2\nu)}, \quad C_{12}^0 = \frac{\nu(1-\nu)}{(1+\nu)(1-2\nu)}, \quad C_{44}^0 = \frac{1}{2}(C_{11}^0 - C_{12}^0), \tag{3.2c}$$

$$C_{11} + 2C_{12} = \frac{1-\nu}{1-2\nu} - \frac{H}{2}. \tag{3.2d}$$

According to $C = C^0 - C^a$, we have the splitting of the operator $\mathcal{D}$, $\mathcal{D} = \mathcal{D}^0 - \mathcal{D}^a$. With this notation, the linear operators $\nabla \cdot \sigma$ and $\sigma \cdot \mathbf{n}$ take the form

$$\begin{aligned} \mathcal{L} &:= \nabla \cdot (\mathcal{D}\mathbf{S}) = \nabla \cdot (\mathcal{D}^0\mathbf{S}) - \nabla \cdot (\mathcal{D}^a\mathbf{S}) = \mathcal{L}^0 - \mathcal{L}^a, \\ \mathcal{B} &:= (\mathcal{D}\mathbf{S}) \cdot \mathbf{n} = (\mathcal{D}^0\mathbf{S}) \cdot \mathbf{n} - (\mathcal{D}^a\mathbf{S}) \cdot \mathbf{n} = \mathcal{B}^0 - \mathcal{B}^a, \end{aligned}$$

and the thermoelastic boundary value problem can be stated in the form

$$\mathcal{L}(\mathbf{w}) = \mathbf{F}, \qquad \mathbf{x} \in \Omega, \ t > 0, \tag{3.3a}$$

$$\mathcal{B}(\mathbf{w}) = \mathbf{g}, \qquad r = R(\phi, z), \tag{3.3b}$$

where

$$\mathbf{F} = (C_{11} + 2C_{12})\nabla\Theta = \left(\frac{1-\nu}{1-2\nu} - \frac{H}{2}\right)\nabla\Theta, \qquad \mathbf{g} = \left(\frac{1-\nu}{1-2\nu} - \frac{H}{2}\right)\Theta\mathbf{n},$$

with $\mathbf{n}$ denoting the outward normal of the surface $r = R(\phi, z)$. The total stress contains an extra diagonal term related to the scaling with respect to the isotropic quantities E and ν so that

$$\sigma_{ij}^{\text{tot,aniso}} = \sigma_{ij} - \left(\frac{1-\nu}{1-2\nu} - \frac{H}{2}\right)\Theta\delta_{ij}. \tag{3.4}$$

We assume that the displacement $\mathbf{w}$ can be written as $\sum_{k=0}^{n} \mathbf{w}_k$. The following procedure defines the $\mathbf{w}_k$ under the assumption that $\mathbf{w}_{k+1} = \mathcal{N}\mathbf{w}_k$ for some linear operator $\mathcal{N}$. To solve for $\mathbf{w}(\mathbf{x})$ in (3.1) we begin by finding the displacement $\mathbf{w}_0$ given by

$$\mathcal{L}^0(\mathbf{w}_0) = \mathbf{F}, \qquad \mathbf{x} \in \Omega, \ t > 0, \tag{3.5a}$$

$$\mathcal{B}^0(\mathbf{w}_0) \ = \mathbf{g}, \qquad r = R(\phi, z), \tag{3.5b}$$

which is the displacement found in [13], multiplied by a factor of $\mu = 1 - \frac{H}{2}\frac{1-2\nu}{1-\nu}$. Having defined $\mathbf{w}_0$, we denote by $\mathbf{w}_{k+1} = \mathcal{N}\mathbf{w}_k$, with $k \geq 0$, the solution to

$$\mathcal{L}^0(\mathbf{w}_{k+1}) = \mathcal{L}^a(\mathbf{w}_k), \qquad \mathbf{x} \in \Omega, \ t > 0, \tag{3.6a}$$

$$\mathcal{B}^0(\mathbf{w}_{k+1}) \ = \mathcal{B}^a(\mathbf{w}_k), \qquad r = R(\phi, z). \tag{3.6b}$$

Continuing this process, we have for $\mathbf{w}(\mathbf{x})$ in (3.1)

$$\mathbf{w} = \mathbf{w}_0 + \mathcal{N}\mathbf{w}_0 + \mathcal{N}^2\mathbf{w}_0 + \cdots + \mathcal{N}^n\mathbf{w}_0 + \cdots . \tag{3.7}$$

Since $\|\mathcal{N}\| \leq \omega$ in a suitable norm, where $\omega = \frac{|H|/2}{C_{11}-C_{12}+H/2} = \frac{|2C_{44}-C_{11}+C_{12}|}{2C_{44}+C_{11}-C_{12}} < 1$ is an anisotropic factor, the series converges and an error can be estimated when (3.7) is replaced by a finite sum; cf. Appendix A. For typical cubic anisotropic materials, $\omega \approx 1/3$ [2].

Vigdergauz and Givoli [11, 12] have discussed the fourfold symmetry case (corresponding to the [001] pulling direction in our case) for a given symmetric temperature field. However, their splitting is not optimal and is valid only for crystals with weak cubic anisotropy. Our procedure can be applied to elastic stress computations with cubic anisotropy under a relatively general setting for a wide variety of materials.

Converting the stress-strain relationship to polar coordinates, we note that C^0 will not change, so we will concern ourselves only with C^a. Corresponding to (3.1) we let $\boldsymbol{\sigma}_{\text{cyc}} = (\sigma_{rr}, \sigma_{\phi\phi}, \sigma_{zz}, \sigma_{\phi z}, \sigma_{rz}, \sigma_{r\phi})^{\text{T}}$, $\mathbf{e}_{\text{cyc}} = (e_{rr}, e_{\phi\phi}, e_{zz}, 2e_{\phi z}, 2e_{rz}, 2e_{r\phi})^{\text{T}}$. The components of C_{cyc} are given by

$$\begin{aligned} C_{\text{cyc},ijkl} = {} & \frac{H}{2}(a_{i1}a_{j1}a_{k1}a_{l1} + a_{i2}a_{j2}a_{k2}a_{l2} + a_{i3}a_{j3}a_{k3}a_{l3}) - \frac{H}{4}(a_{i2}a_{j3}a_{k2}a_{l3} \\ & + a_{i2}a_{j3}a_{k3}a_{l2} + a_{i3}a_{j2}a_{k2}a_{l3} + a_{i3}a_{j2}a_{k3}a_{l2} + a_{i1}a_{j3}a_{k1}a_{l3} + a_{i1}a_{j3}a_{k3}a_{l1} \\ & + a_{i3}a_{j1}a_{k1}a_{l3} + a_{i3}a_{j1}a_{k3}a_{l1} + a_{i1}a_{j2}a_{k1}a_{l2} + a_{i1}a_{j2}a_{k2}a_{l1} + a_{i2}a_{j1}a_{k1}a_{l2} \\ & + a_{i2}a_{j1}a_{k2}a_{l1}) \end{aligned}$$

with a_{ij} the cosine of the angle between x'_i (new axes) and x_j (old axes) [10]. Furthermore, the first two and last two suffixes are abbreviated into a single suffix according to the scheme $11 \to 1$; $22 \to 2$; $33 \to 3$; $23, 32 \to 4$; $13, 31 \to 5$; $12, 21 \to 6$. For example, $C_{\text{cyc},1111} \equiv C_{\text{cyc},11}$ and $C_{\text{cyc},1231} \equiv C_{\text{cyc},65}$.

For the [001] pulling direction, we choose the z-direction as [001], and the directions [100] and [010] correspond to $\phi = 0$ and $\phi = \pi/2$, respectively, so that

$$a_{ij}^{[001]} = \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For the $[\bar{1}\bar{1}\bar{1}]$ pulling direction, the z-direction is $[\bar{1}\bar{1}\bar{1}]$ and we choose $\phi = 0$ and $\phi = \pi/2$ to correspond to $[2\bar{1}\bar{1}]$ and $[0\bar{1}1]$, respectively. In this case,

$$a_{ij}^{[\bar{1}\bar{1}\bar{1}]} = \begin{pmatrix} \frac{2}{\sqrt{6}}\cos\phi & -\frac{1}{\sqrt{6}}\cos\phi - \frac{1}{\sqrt{2}}\sin\phi & -\frac{1}{\sqrt{6}}\cos\phi + \frac{1}{\sqrt{2}}\sin\phi \\ -\frac{2}{\sqrt{6}}\sin\phi & \frac{1}{\sqrt{6}}\sin\phi - \frac{1}{\sqrt{2}}\cos\phi & \frac{1}{\sqrt{6}}\sin\phi + \frac{1}{\sqrt{2}}\cos\phi \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{pmatrix}.$$

Finally, for the $[\bar{2}11]$ pulling direction, the z-direction is $[\bar{2}11]$, $\phi = 0$ corresponds to [111], and $\phi = \pi/2$ corresponds to $[01\bar{1}]$ yielding

$$a_{ij}^{[\bar{2}11]} = \begin{pmatrix} \frac{1}{\sqrt{3}}\cos\phi & \frac{1}{\sqrt{3}}\cos\phi + \frac{1}{\sqrt{2}}\sin\phi & \frac{1}{\sqrt{3}}\cos\phi - \frac{1}{\sqrt{2}}\sin\phi \\ -\frac{1}{\sqrt{3}}\sin\phi & -\frac{1}{\sqrt{3}}\sin\phi + \frac{1}{\sqrt{2}}\cos\phi & -\frac{1}{\sqrt{3}}\sin\phi - \frac{1}{\sqrt{2}}\cos\phi \\ -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

Each of these transformations changes the form of C_{cyc}, and in general we have the form $C_{\text{cyc}} = \sum_k C_{\text{cyc,k}}$, where each matrix with subscript k consists of only elements $c_k = \cos(k\phi)$, $s_k = \sin(k\phi)$ and zero. The detailed expressions are given in Appendix B. Since the anisotropic part of the constitutive relation C^a_{cyc} can be decomposed into components $C^a_{\text{cyc},k}$ consisting of only s_k and c_k, we can systematically construct higher order approximations. This is accomplished by first determining the solution for a generic $C^a_{\text{cyc},k}$, and then computing an appropriate linear combination of all the solutions for a particular pulling direction. To illustrate the procedure, we now discuss a simpler problem, where we use the plane-strain assumption.

3.2. Plane-strain thermal stress for solids with cubic anisotropy. As in [13], we assume that the displacement is only in the (r, ϕ) plane so that $\mathbf{e}_{\mathrm{cyc}} = (e_{rr}, e_{\phi\phi}, 0, 0, 0, 2e_{r\phi})^{\mathrm{T}}$. We will also reintroduce the notation $(\mathcal{C}_k, \mathcal{S}_k) = (\cos(n_k\phi + \delta_k), \sin(n_k\phi + \delta_k))$ and its generalizations $(\mathcal{C}, \mathcal{S})$ and $(\tilde{\mathcal{C}}_k, \tilde{\mathcal{S}}_k)$ with the k suffix suppressed, and $\tilde{n}_k$ replacing n_k, respectively.

Starting with the isotropic case, where $\mathbf{w}_0$ is the solution of (3.5), when ϵ is small it is shown in [13] that $\mathbf{w}_0$ can be approximated by

$$\mathbf{w}_0 \sim \begin{pmatrix} rD_r^{(1)} + r^3 D_r^{(3)} \\ 0 \end{pmatrix} + r^{n_k - 1} \begin{pmatrix} D_r^- \mathcal{C}_k \\ D_\phi^- \mathcal{S}_k \end{pmatrix} + r^{n_k + 1} \begin{pmatrix} D_r^+ \mathcal{C}_k \\ D_\phi^+ \mathcal{S}_k \end{pmatrix}, \tag{3.8}$$

where

$$D_r^{(1)} = \mu \left(\frac{1+\nu}{1-\nu}\right) C_1(1-2\nu), \qquad D_r^{(3)} = \mu \left(\frac{1+\nu}{1-\nu}\right) \frac{C_1}{\bar{R}^2}, \tag{3.9a}$$

$$D_r^- = \mu \left(\frac{1+\nu}{1-\nu}\right) \frac{C_1 \epsilon \beta_k}{\bar{R}^{n_k - 2}} n_k, \qquad D_\phi^- = -\mu \left(\frac{1+\nu}{1-\nu}\right) \frac{C_1 \epsilon \beta_k}{\bar{R}^{n_k - 2}} n_k, \tag{3.9b}$$

$$D_r^+ = \mu \left(\frac{1+\nu}{1-\nu}\right) \frac{C_1 \epsilon \beta_k}{\bar{R}^{n_k}} (2 - 4\nu - n_k) + \frac{4\mu(1+\nu)}{n_k(n_k+1)} \frac{C_2 \epsilon \beta_k}{\bar{R}^{n_k}}, \tag{3.9c}$$

$$D_\phi^+ = \mu \left(\frac{1+\nu}{1-\nu}\right) \frac{C_1 \epsilon \beta_k}{\bar{R}^{n_k}} (4 - 4\nu + n_k) + \frac{4\mu(1+\nu)}{n_k(n_k+1)} \frac{C_2 \epsilon \beta_k}{\bar{R}^{n_k}}, \tag{3.9d}$$

where $\mu = 1 - \frac{H}{2} \frac{1-2\nu}{1-\nu}$ (see section 3.1).

Having determined $\mathbf{w}_0$, $\mathbf{w}$ is given by the expansion (3.7). Each of the terms in the expansion is a solution of the boundary value problem (3.6). To illustrate the procedure, in the following we construct $\mathbf{w}_1 = \mathcal{N}\mathbf{w}_0$.

Due to the linearity of the equilibrium equation, we can pick a representative $\mathbf{v} = (D_r r^k \cos(n\phi + \delta), D_\phi r^k \sin(n\phi + \delta))^{\mathrm{T}}$ with $n \geq 0$, $k \geq 1$. From this $\mathbf{v}$, the strain

$$\mathbf{S}(\mathbf{v}) = \begin{pmatrix} e_{rr} \\ e_{\phi\phi} \\ 2e_{r\phi} \end{pmatrix} = \begin{pmatrix} kD_r r^{k-1} \mathcal{C} \\ (D_r + nD_\phi) r^{k-1} \mathcal{C} \\ (kD_\phi - D_\phi - nD_r) r^{k-1} \mathcal{S} \end{pmatrix},$$

and the stress due to the anisotropy in the material parameters is given by $C_{\mathrm{cyc}}^a \mathbf{S}(\mathbf{v})$, where the exact form of C_{cyc}^a depends on the orientation of the crystal. Expressions (B.1)–(B.3) show that C_{cyc}^a is a sum of terms, $C_{\mathrm{cyc},m}^a$, characterized by $\cos m\phi$ and $\sin m\phi$. Therefore, $C_{\mathrm{cyc},m}^a \mathbf{S}(\mathbf{v})$ can be expressed as a sum with terms of the form $r^{k-1}(\cos(\tilde{n}\phi + \delta), \cos(\tilde{n}\phi + \delta), \sin(\tilde{n}\phi + \delta))^{\mathrm{T}}$, $\tilde{n} = n \pm m$. So, we need only consider the problem

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{1}{r} \frac{\partial \sigma_{r\phi}}{\partial \phi} + \frac{\sigma_{rr} - \sigma_{\phi\phi}}{r} = f_r r^{k-2} \tilde{\mathcal{C}}, \qquad r < \bar{R}(z), \tag{3.10a}$$

$$\frac{\partial \sigma_{r\phi}}{\partial r} + \frac{1}{r} \frac{\partial \sigma_{\phi\phi}}{\partial \phi} + \frac{2\sigma_{r\phi}}{r} = f_\phi r^{k-2} \tilde{\mathcal{S}}, \qquad r < \bar{R}(z), \tag{3.10b}$$

with integers $\tilde{n} \geq 0$, $k \geq 1$, and

$$\sigma_{rr} = g_r r^{k-1} \tilde{\mathcal{C}}, \qquad r = \bar{R}(z), \tag{3.11a}$$

$$\sigma_{r\phi} = g_\phi r^{k-1} \tilde{\mathcal{S}}, \qquad r = \bar{R}(z), \tag{3.11b}$$

corresponding to (3.6) with the higher order terms omitted.

To determine the solution to (3.10)–(3.11) we take a two-step approach. We begin by finding a particular solution $\mathbf{w}_p$ which satisfies (3.10) but not necessarily (3.11). Next, we find $\mathbf{w}_h$ which solves the homogeneous version of (3.10) and the modified boundary condition

$$\sigma^h_{rr} = g_r r^{k-1}\tilde{\mathcal{C}} - \sigma^p_{rr} := \tilde{g}_r r^{k-1}\tilde{\mathcal{C}}, \qquad r = \bar{R}(z), \tag{3.12a}$$

$$\sigma^h_{r\phi} = g_\phi r^{k-1}\tilde{\mathcal{S}} - \sigma^p_{r\phi} := \tilde{g}_\phi r^{k-1}\tilde{\mathcal{S}}, \qquad r = \bar{R}(z), \tag{3.12b}$$

where σ^p_{rr} and $\sigma^p_{r\phi}$ are stress components corresponding to $\mathbf{w}_p$. Accordingly, we find

$$\mathbf{w}_p = \begin{cases} \frac{1+\nu}{(1-\nu)^2}\begin{pmatrix} a_r r^k\tilde{\mathcal{C}} \\ a_\phi r^k \tilde{\mathcal{S}}\end{pmatrix}, & (k\pm\tilde{n})^2 \neq 1, \\ \frac{1+\nu}{(1-\nu)^2}\begin{pmatrix} (b_r + b_\phi\zeta\ln r)r^k\tilde{\mathcal{C}} \\ (b_\phi r^k \ln r\tilde{\mathcal{S}})\end{pmatrix}, & (k\pm\tilde{n})^2 = 1, \end{cases} \tag{3.13a}$$

where

$$a_r = \frac{((1-2\nu)(k^2-1) - 2\tilde{n}^2(1-\nu))f_r + \tilde{n}(3-4\nu-k)f_\phi}{((k-\tilde{n})^2-1)((k+\tilde{n})^2-1)}, \tag{3.13b}$$

$$a_\phi = \frac{\tilde{n}(3-4\nu+k)f_r + (2(1-\nu)(k^2-1) - (1-2\nu)\tilde{n}^2)f_\phi}{((k-\tilde{n})^2-1)((k+\tilde{n})^2-1)}, \tag{3.13c}$$

$$b_r = \begin{cases} -\frac{(3-4\nu)(\tilde{n}-1)(f_r+f_\phi)+(f_r-f_\phi)}{8\tilde{n}(\tilde{n}-1)}, & k = \tilde{n}-1, \\ \frac{(3-4\nu)\tilde{n}^2(f_r+f_\phi)+8(1-\nu)(1-2\nu)(\tilde{n}+1)(f_r-f_\phi)}{8\tilde{n}(\tilde{n}+1)(\tilde{n}+4-4\nu)}, & k = \tilde{n}+1, \end{cases} \tag{3.13d}$$

$$b_\phi = \begin{cases} -\frac{(\tilde{n}-1)(f_r+f_\phi)+(3-4\nu)(f_r-f_\phi)}{8(\tilde{n}-1)}, & k = \tilde{n}-1, \\ \frac{(\tilde{n}+4-4\nu)(f_r+f_\phi)}{8(\tilde{n}+1)}, & k = \tilde{n}+1, \end{cases} \tag{3.13e}$$

$$\zeta = \begin{cases} -1, & k = \tilde{n}-1, \\ -\frac{\tilde{n}-2+4\nu}{\tilde{n}+4-4\nu}, & k = \tilde{n}+1. \end{cases} \tag{3.13f}$$

The special case when $\tilde{n} = 0$ and $k = 1$ takes the form $\mathbf{w}_p = \frac{1+\nu}{2(1-\nu)^2} r\ln r \times ((1-2\nu)f_r\cos\delta,\, 2(1-\nu)f_\phi\sin\delta)^{\mathrm{T}}$. Corresponding to $\mathbf{w}_p$ are the stress components

$$\begin{pmatrix} \sigma^p_{rr} \\ \sigma^p_{\phi\phi} \\ \sigma^p_{r\phi} \end{pmatrix} = \begin{pmatrix} \frac{1-\nu}{(1+\nu)(1-2\nu)}((k-k\nu+\nu)a_r + \nu\tilde{n}a_\phi)r^{k-1}\tilde{\mathcal{C}} \\ \frac{1-\nu}{(1+\nu)(1-2\nu)}((1-\nu+k\nu)a_r + (1-\nu)\tilde{n}a_\phi)r^{k-1}\tilde{\mathcal{C}} \\ \frac{1-\nu}{2(1+\nu)}(-\tilde{n}a_r + (k-1)a_\phi)r^{k-1}\tilde{\mathcal{S}} \end{pmatrix} \tag{3.14a}$$

for $(k\pm\tilde{n})^2 \neq 1$ and

$$\begin{pmatrix} \sigma^p_{rr} \\ \sigma^p_{\phi\phi} \\ \sigma^p_{r\phi} \end{pmatrix} = \begin{pmatrix} \frac{1}{(1-\nu)(1-2\nu)}c_{rr}r^{k-1}\tilde{\mathcal{C}} \\ \frac{1}{(1-\nu)(1-2\nu)}c_{\phi\phi}r^{k-1}\tilde{\mathcal{C}} \\ \frac{1}{2(1-\nu)}c_{r\phi}r^{k-1}\tilde{\mathcal{S}} \end{pmatrix} \tag{3.14b}$$

for $(k \pm \tilde{n})^2 = 1$, where

$$c_{rr} = (k - k\nu + \nu)(b_r + b_\phi \zeta \ln r) + b_\phi((1-\nu)\zeta + \nu \tilde{n} \ln(r)), \tag{3.14c}$$

$$c_{\phi\phi} = (1 - \nu + k\nu)(b_r + b_\phi \zeta \ln r) + b_\phi(\nu\zeta + (1-\nu)\tilde{n} \ln(r)), \tag{3.14d}$$

$$c_{r\phi} = -\tilde{n}(b_r + b_\phi \zeta \ln r) + b_\phi(1 + (k-1)\ln r). \tag{3.14e}$$

For the special case when $\tilde{n} = 0$ and $k = 1$, we have

$$(\sigma^p_{rr}, \sigma^p_{\phi\phi}, \sigma^p_{r\phi})^{\mathrm{T}} = \frac{1}{2(1-\nu)}(f_r(\ln r + 1 - \nu)\cos\delta, f_r(\ln r + \nu)\cos\delta, f_\phi(1-\nu)\sin\delta)^{\mathrm{T}}.$$

Using the technique described in [13], we can find $\mathbf{w}_h$ which solves the homogeneous version of (3.10) and the boundary condition (3.12),

$$\mathbf{w}_h = \frac{1+\nu}{2(1-\nu)} \begin{pmatrix} \left(\frac{(2-\tilde{n}-4\nu)(\tilde{g}_r+\tilde{g}_\phi)r^{\tilde{n}+1}}{(\tilde{n}+1)\bar{R}^{\tilde{n}}} + \frac{(\tilde{n}\tilde{g}_r+(\tilde{n}-2)\tilde{g}_\phi)r^{\tilde{n}-1}}{(\tilde{n}-1)\bar{R}^{\tilde{n}-2}}\right)\tilde{\mathcal{C}} \\ \left(\frac{(4+\tilde{n}-4\nu)(\tilde{g}_r+\tilde{g}_\phi)r^{\tilde{n}+1}}{(\tilde{n}+1)\bar{R}^{\tilde{n}}} - \frac{(\tilde{n}\tilde{g}_r+(\tilde{n}-2)\tilde{g}_\phi)r^{\tilde{n}-1}}{(\tilde{n}-1)\bar{R}^{\tilde{n}-2}}\right)\tilde{\mathcal{S}} \end{pmatrix}. \tag{3.15}$$

The corresponding stress components are given by

$$\begin{pmatrix} \sigma^h_{rr} \\ \sigma^h_{\phi\phi} \\ \sigma^h_{r\phi} \end{pmatrix} = \begin{pmatrix} \left(\frac{(2-\tilde{n})(\tilde{g}_r+\tilde{g}_\phi)r^{\tilde{n}}}{2\bar{R}^{\tilde{n}}} + \frac{(\tilde{n}\tilde{g}_r+(\tilde{n}-2)\tilde{g}_\phi)r^{\tilde{n}-2}}{2\bar{R}^{\tilde{n}-2}}\right)\tilde{\mathcal{C}} \\ \left(\frac{(2+\tilde{n})(\tilde{g}_r+\tilde{g}_\phi)r^{\tilde{n}}}{2\bar{R}^{\tilde{n}}} - \frac{(\tilde{n}\tilde{g}_r+(\tilde{n}-2)\tilde{g}_\phi)r^{\tilde{n}-2}}{2\bar{R}^{\tilde{n}-2}}\right)\tilde{\mathcal{C}} \\ \left(\frac{\tilde{n}(\tilde{g}_r+\tilde{g}_\phi)r^{\tilde{n}}}{2\bar{R}^{\tilde{n}}} - \frac{(\tilde{n}\tilde{g}_r+(\tilde{n}-2)\tilde{g}_\phi)r^{\tilde{n}-2}}{2\bar{R}^{\tilde{n}-2}}\right)\tilde{\mathcal{S}} \end{pmatrix}. \tag{3.16}$$

In the special case when $\tilde{n} = 0$ (or $\tilde{n} = 1$), we require $\tilde{g}_\phi = 0$ (or $\tilde{g}_r = \tilde{g}_\phi$) for the homogeneous elasticity problem to be well-posed. The solution and the corresponding stress components are given by (3.15) and (3.16) without the term related to $r^{\tilde{n}-1}$ and $r^{\tilde{n}-2}$, respectively.

In the following we find the explicit form of the expression $\mathbf{w}_1 = \mathcal{N}\mathbf{w}_0$ for the $[\bar{1}\bar{1}\bar{1}]$ pulling direction. This expression generates the first order corrections to the stress of an anisotropic cubic crystal. The outline of the procedure is also given for the [001] and $[\bar{2}11]$ seeding orientations.

3.2.1. $[\bar{1}\bar{1}\bar{1}]$ pulling direction. To treat this case systematically we decompose $\mathbf{w}_0$ into five separate quantities given by

$$\mathbf{w}_{0,A} = \begin{pmatrix} D^{(1)}_r r \\ 0 \end{pmatrix}, \quad \mathbf{w}_{0,B} = \begin{pmatrix} D^{(3)}_r r^3 \\ 0 \end{pmatrix}, \quad \mathbf{w}_{0,C} = D^-_r r^k \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix},$$
$$\mathbf{w}_{0,D} = \frac{D^+_r + D^+_\phi}{2} r^k \begin{pmatrix} \mathcal{C}_k \\ \mathcal{S}_k \end{pmatrix}, \quad \mathbf{w}_{0,E} = \frac{D^+_r - D^+_\phi}{2} r^k \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix}.$$

For $\mathbf{w}_{0,C}$, $k = n_k - 1$, while for both $\mathbf{w}_{0,D}$ and $\mathbf{w}_{0,E}$, $k = n_k + 1$. What characterizes the $[\bar{1}\bar{1}\bar{1}]$ direction is the anisotropic stiffness given by C^a. From (B.1), we have in the case of plane-strain $C^a = C^a_0$, where

$$C^{a,[\bar{1}\bar{1}\bar{1}]} = \frac{H}{12} \begin{pmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

and as a result $\tilde{n} = n$ for the $[\bar{1}\bar{1}\bar{1}]$ direction.

For the first component, $\mathbf{w}_{0,A}$, we find from (3.6a) and (3.10) that

$$\mathcal{L}^a(\mathbf{w}_{0,A}) = \nabla \cdot C^{a,[\bar{1}\bar{1}\bar{1}]}\mathbf{S}(\mathbf{w}_{0,A}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = r^{k-2}\begin{pmatrix} f_r\mathcal{C} \\ f_\phi\mathcal{S} \end{pmatrix}, \tag{3.17a}$$

and for the boundary condition, (3.6b) and (3.11) give

$$\mathcal{B}^a(\mathbf{w}_{0,A}) = C^{a,[\bar{1}\bar{1}\bar{1}]}\mathbf{S}(\mathbf{w}_{0,A}) \cdot \mathbf{n} = \frac{H}{6}\begin{pmatrix} D_r^{(1)} \\ 0 \end{pmatrix} = \bar{R}^{k-1}\begin{pmatrix} g_r\mathcal{C} \\ g_\phi\mathcal{S} \end{pmatrix}, \tag{3.17b}$$

where $k = 1$, $\delta = 0$, and $n = \tilde{n} = n_k = 0$. Setting $\Lambda_A = \frac{H}{6}D_r^{(1)}$ we identify $f_r = f_\phi = 0$, $g_r = \Lambda_A$, and $g_\phi = 0$. The quantities f_r and f_ϕ applied to (3.13)–(3.14) give the particular solution for the stress as $\sigma^p_{ij,A} = 0$, which through (3.12) indicates that $\tilde{g}_r = \Lambda_A$ and $\tilde{g}_\phi = 0$. For the homogeneous solution, we solve $\mathcal{L}^0(\mathbf{w}_{1,A}) = \mathcal{L}^a(\mathbf{w}_{0,A})$ with the boundary condition $\mathcal{B}^0(\mathbf{w}_{1,A}) = \mathcal{B}^a(\mathbf{w}_{0,A})$ and by using (3.16) to determine the stress, which gives

$$\begin{pmatrix} \sigma^h_{rr,A} \\ \sigma^h_{\phi\phi,A} \\ \sigma^h_{r\phi,A} \end{pmatrix} = \Lambda_A \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}. \tag{3.17c}$$

For $\mathbf{w}_{0,B}$, we have $k = 3$, $\delta = 0$, and $n = \tilde{n} = n_k = 0$, and continuing in an analogous fashion we find that

$$\mathcal{L}^a(\mathbf{w}_{0,B}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathcal{B}^a(\mathbf{w}_{0,B}) = \bar{R}^2\begin{pmatrix} g_r \\ g_\phi \end{pmatrix} = \frac{H}{6}\bar{R}^2\begin{pmatrix} D_r^{(3)} \\ 0 \end{pmatrix}. \tag{3.18a}$$

As with the previous case, we find $\sigma^p_{ij,B} = 0$, and defining $\Lambda_B = \frac{H}{6}D_r^{(3)}$ we identify $\tilde{g}_r = g_r = \Lambda_B$, and $\tilde{g}_\phi = g_\phi = f_r = f_\phi = 0$. From (3.16) we have

$$\begin{pmatrix} \sigma^h_{rr,B} \\ \sigma^h_{\phi\phi,B} \\ \sigma^h_{r\phi,B} \end{pmatrix} = \Lambda_B\bar{R}^2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}. \tag{3.18b}$$

For $\mathbf{w}_{0,C}$, $k = n_k - 1$, $n = \tilde{n} = n_k$, and defining $\Lambda_C = \frac{H}{6}(1 - n_k)D_r^-$ we determine

$$\mathcal{L}^a(\mathbf{w}_{0,C}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathcal{B}^a(\mathbf{w}_{0,C}) = \Lambda_C\bar{R}^{n_k-2}\begin{pmatrix} \mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix}$$

so that $f_r = f_\phi = 0$, $\sigma^p_{ij,C} = 0$, $g_r = \tilde{g}_r = \Lambda_C$, and $g_\phi = \tilde{g}_\phi = -\Lambda_C$. Applying (3.16) we obtain

$$\begin{pmatrix} \sigma^h_{rr,C} \\ \sigma^h_{\phi\phi,C} \\ \sigma^h_{r\phi,C} \end{pmatrix} = \Lambda_C r^{n_k-2} \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix}. \tag{3.19}$$

The fourth component is $\mathbf{w}_{0,D}$ and $k = n_k + 1$, $n = \tilde{n} = n_k$. By defining $\Lambda_D = \frac{H}{12}(n_k + 1)(D_r^+ + D_\phi^+)$ one has

$$\mathcal{L}^a(\mathbf{w}_{0,D}) = n_k \Lambda_D r^{n_k - 1} \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix}, \qquad \mathcal{B}^a(\mathbf{w}_{0,D}) = \Lambda_D \bar{R}^{n_k} \begin{pmatrix} \mathcal{C}_k \\ 0 \end{pmatrix}$$

so that $f_r = -f_\phi = n_k \Lambda_D$, $g_r = \Lambda_D$, and $g_\phi = 0$. In this case the particular solution for the stress is

$$\begin{pmatrix} \sigma^p_{rr,D} \\ \sigma^p_{\phi\phi,D} \\ \sigma^p_{r\phi,D} \end{pmatrix} = \frac{\Lambda_D r^{n_k}}{n_k + 4 - 4\nu} \begin{pmatrix} 2(n_k + 1 - \nu n_k)\mathcal{C}_k \\ 2(1 + \nu n_k)\mathcal{C}_k \\ -n_k(1 - 2\nu)\mathcal{S}_k \end{pmatrix} \tag{3.20a}$$

so that

$$\begin{pmatrix} \tilde{g}_r \\ \tilde{g}_\phi \end{pmatrix} = \frac{(1 - 2\nu)\Lambda_D}{n_k + 4 - 4\nu} \begin{pmatrix} 2 - n_k \\ n_k \end{pmatrix},$$

and from (3.16)

$$\begin{pmatrix} \sigma^h_{rr,D} \\ \sigma^h_{\phi\phi,D} \\ \sigma^h_{r\phi,D} \end{pmatrix} = \frac{(1 - 2\nu)\Lambda_D r^{n_k}}{n_k + 4 - 4\nu} \begin{pmatrix} (2 - n_k)\mathcal{C}_k \\ (n_k + 2)\mathcal{C}_k \\ n_k \mathcal{S}_k \end{pmatrix}. \tag{3.20b}$$

The last component, $\mathbf{w}_{0,E}$, has $k = n_k + 1$ and $n = \tilde{n} = n_k$. For this case we choose $\Lambda_E = \frac{H}{12}(D_\phi^+ - D_r^+)$ and find

$$\mathcal{L}^a(\mathbf{w}_{0,E}) = n_k \Lambda_E r^{n_k - 1} \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix}, \qquad \mathcal{B}^a(\mathbf{w}_{0,E}) = \Lambda_E \bar{R}^{n_k} \begin{pmatrix} (n_k - 1)\mathcal{C}_k \\ -n_k \mathcal{S}_k \end{pmatrix}$$

so that $f_r = -f_\phi = n_k \Lambda_E$, $g_r = (n_k - 1)\Lambda_E$, and $g_\phi = -n_k \Lambda_E$. Continuing,

$$\begin{pmatrix} \sigma^p_{rr,E} \\ \sigma^p_{\phi\phi,E} \\ \sigma^p_{r\phi,E} \end{pmatrix} = \frac{\Lambda_E r^{n_k}}{n_k + 4 - 4\nu} \begin{pmatrix} 2(n_k + 1 - \nu n_k)\mathcal{C}_k \\ 2(1 + \nu n_k)\mathcal{C}_k \\ -n_k(1 - 2\nu)\mathcal{S}_k \end{pmatrix}, \tag{3.21a}$$

yielding

$$\begin{pmatrix} \tilde{g}_r \\ \tilde{g}_\phi \end{pmatrix} = \frac{(n_k + 3 - 2\nu)\Lambda_E}{n_k + 4 - 4\nu} \begin{pmatrix} n_k - 2 \\ -n_k \end{pmatrix}$$

and

$$\begin{pmatrix} \sigma^h_{rr,E} \\ \sigma^h_{\phi\phi,E} \\ \sigma^h_{r\phi,E} \end{pmatrix} = -\frac{(n_k + 3 - 2\nu)\Lambda_E r^{n_k}}{n_k + 4 - 4\nu} \begin{pmatrix} (2 - n_k)\mathcal{C}_k \\ (n_k + 2)\mathcal{C}_k \\ n_k \mathcal{S}_k \end{pmatrix}. \tag{3.21b}$$

Combining (3.17)–(3.21) and using both (3.9) and (3.4), we find the first order correction to the total stress in the $[\bar{1}\bar{1}\bar{1}]$ direction accounting for cubic anisotropy as

$$\begin{aligned}
\begin{pmatrix} \sigma_{rr}^{\mathrm{tot}} \\ \sigma_{\phi\phi}^{\mathrm{tot}} \\ \sigma_{zz}^{\mathrm{tot}} \\ \sigma_{r\phi}^{\mathrm{tot}} \end{pmatrix}_{[\bar{1}\bar{1}\bar{1}]}
&= \frac{2(1-\nu)\omega C_1}{3} \begin{pmatrix} 1 \\ 1 \\ 2\nu \\ 0 \end{pmatrix} - \frac{4(1-\nu)^2 \omega C_1 r^2}{(1+\nu)(1-2\nu)\bar{R}^2} \begin{pmatrix} 1 \\ 1 \\ 1+\nu \\ 0 \end{pmatrix} \\
&\quad - \frac{4(1-\nu)\epsilon\omega C_2}{3(1-2\nu)} \frac{\beta_k}{n_k} \left(\frac{r}{\bar{R}}\right)^{n_k} \mathcal{C}_k \begin{pmatrix} 2\frac{1-\nu+\nu^2}{1+\nu} \\ 2\frac{1-\nu+\nu^2}{1+\nu} \\ 3-5\nu+4\nu^2 \\ 0 \end{pmatrix} \\
&\quad + \frac{\epsilon\omega C_1}{3} \beta_k (n_k+1) \left(\frac{r}{\bar{R}}\right)^{n_k} \begin{pmatrix} (n_k+2-4\nu)\mathcal{C}_k \\ (2-n_k-4\nu)\mathcal{C}_k \\ 4\nu(1-2\nu)\mathcal{C}_k \\ -n_k\mathcal{S}_k \end{pmatrix} \\
&\quad - \frac{\epsilon\omega C_1}{3} \beta_k n_k (n_k-1) \left(\frac{r}{\bar{R}}\right)^{n_k-2} \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{C}_k \\ 0 \\ -\mathcal{S}_k \end{pmatrix}
\end{aligned} \tag{3.22}$$

with $\omega = \frac{1+\nu}{1-\nu}\frac{|H|}{2}$ using the scaled version of H.

This procedure can of course be followed for any pulling direction provided the form of C^a is known. It can also be applied to finding higher order corrections provided that the solution (3.13) to (3.10) is generalized to allow a multiplicative factor of $(\ln r)^l$ for some integer $l \geq 1$. In the following we simply state $\mathcal{L}^a$ and $\mathcal{B}^a$ for the [001] and $[\bar{2}11]$ seeding orientations for each of the five components of the displacement (3.8).

3.2.2. [001] pulling direction. From (B.2) we have $C^a = C_0^a + C_4^a$ with

$$C_0^a = \frac{H}{4}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad C_4^a = \frac{H}{4}\begin{pmatrix} c_4 & -c_4 & -s_4 \\ -c_4 & c_4 & s_4 \\ -s_4 & s_4 & -c_4 \end{pmatrix}.$$

Accordingly one finds that

$$\begin{aligned}
\mathcal{L}^a(\mathbf{w}_{0,A}) &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & \mathcal{B}^a(\mathbf{w}_{0,A}) &= 3\Lambda_A \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
\mathcal{L}^a(\mathbf{w}_{0,B}) &= 12\Lambda_B r \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & \mathcal{B}^a(\mathbf{w}_{0,B}), &= 3\Lambda_B \bar{R}^2 \begin{pmatrix} 2+c_4 \\ -s_4 \end{pmatrix}.
\end{aligned}$$

The composite form of $\mathcal{B}^a(\mathbf{w}_{0,B})$ shows that the condition (3.11) can generate more than one term of a particular solution for a fixed version of (3.10). For the rest of the terms we extend the notation $\{\mathcal{C}_k, \mathcal{S}_k\}$ to $\{\mathcal{C}_{k,m}, \mathcal{S}_{k,m}\}$, where $\mathcal{C}_{k,m} = \cos((n_k - m)\phi + \delta_k)$, and $\mathcal{S}_{k,m}$ is updated similarly. In this notation, one has

$$\mathcal{L}^a(\mathbf{w}_{0,C}) = 6(2-n_k)\Lambda_C r^{n_k-3}\begin{pmatrix}\mathcal{C}_{k,4}\\ \mathcal{S}_{k,4}\end{pmatrix}, \quad \mathcal{B}^a(\mathbf{w}_{0,C}) = -3\Lambda_C \bar{R}^{n_k-2}\begin{pmatrix}\mathcal{C}_{k,4}\\ \mathcal{S}_{k,4}\end{pmatrix},$$

$$\mathcal{L}^a(\mathbf{w}_{0,D}) = 3n_k\Lambda_D r^{n_k-1}\begin{pmatrix}\mathcal{C}_{k}\\ -\mathcal{S}_{k}\end{pmatrix}, \quad \mathcal{B}^a(\mathbf{w}_{0,D}) = 3\Lambda_D \bar{R}^{n_k}\begin{pmatrix}\mathcal{C}_{k}\\ 0\end{pmatrix},$$

$$\mathcal{L}^a(\mathbf{w}_{0,E}) = 3n_k\Lambda_E r^{n_k-1}\begin{pmatrix}2(1-n_k)\mathcal{C}_{k,4} - \mathcal{C}_k\\ 2(1-n_k)\mathcal{S}_{k,4} + \mathcal{S}_k\end{pmatrix},$$

$$\mathcal{B}^a(\mathbf{w}_{0,E}) = -3\Lambda_E \bar{R}^{n_k}\begin{pmatrix}n_k\mathcal{C}_{k,4} + \mathcal{C}_k\\ n_k\mathcal{S}_{k,4}\end{pmatrix}.$$

3.2.3. $[\bar{2}11]$ pulling direction. From (B.3),

$$C^{a,[\bar{2}11]} = \frac{H}{48}\begin{pmatrix}3-4c_2-7c_4 & 9+7c_4 & 2s_2+7s_4\\ 9+7c_4 & 3+4c_2-7c_4 & 2s_2-7s_4\\ 2s_2+7s_4 & 2s_2-7s_4 & -3+7c_4\end{pmatrix}.$$

In the case for $[\bar{2}11]$, C^a is decoupled into C^a_0, C^a_2, C^a_4, which is analogous to the [001] case. Repeating the calculation we find

$$\begin{aligned}
\mathcal{L}^a(\mathbf{w}_{0,A}) &= \begin{pmatrix}0\\0\end{pmatrix},\\
\mathcal{B}^a(\mathbf{w}_{0,A}) &= \frac{1}{2}\Lambda_A\begin{pmatrix}3-c_2\\ s_2\end{pmatrix},\\
\mathcal{L}^a(\mathbf{w}_{0,B}) &= 3\Lambda_B r\begin{pmatrix}1-c_2\\ s_2\end{pmatrix},\\
\mathcal{B}^a(\mathbf{w}_{0,B}) &= \frac{1}{4}\Lambda_B\bar{R}^2\begin{pmatrix}-7c_4-6c_2+9\\ 7s_4+4s_2\end{pmatrix},\\
\mathcal{L}^a(\mathbf{w}_{0,C}) &= \frac{1}{2}(n_k-2)\Lambda_C r^{n_k-3}\begin{pmatrix}7\mathcal{C}_{k,4}+\mathcal{C}_{k,2}\\ 7\mathcal{S}_{k,4}-\mathcal{S}_{k,2}\end{pmatrix},\\
\mathcal{B}^a(\mathbf{w}_{0,C}) &= \frac{1}{4}\Lambda_C\bar{R}^{n_k-2}\begin{pmatrix}7\mathcal{C}_{k,4}+2\mathcal{C}_{k,2}+3\mathcal{C}_k\\ 7\mathcal{S}_{k,4}-3\mathcal{S}_k\end{pmatrix},\\
\mathcal{L}^a(\mathbf{w}_{0,D}) &= \frac{1}{2}n_k\Lambda_D r^{n_k-1}\begin{pmatrix}-\mathcal{C}_{k,2}+3\mathcal{C}_k\\ -\mathcal{S}_{k,2}-3\mathcal{S}_k\end{pmatrix},\\
\mathcal{B}^a(\mathbf{w}_{0,D}) &= \frac{1}{4}\Lambda_D\bar{R}^{n_k}\begin{pmatrix}-\mathcal{C}_{k,2}-\mathcal{C}_{k,-2}+6\mathcal{C}_k\\ -\mathcal{S}_{k,2}+\mathcal{S}_{k,-2}\end{pmatrix},\\
\mathcal{L}^a(\mathbf{w}_{0,E}) &= \frac{1}{2}n_k\Lambda_E r^{n_k-1}\begin{pmatrix}7(n_k-1)\mathcal{C}_{k,4}+(n_k+1)\mathcal{C}_{k,2}\\ 7(n_k-1)\mathcal{S}_{k,4}+(3-n_k)\mathcal{S}_{k,2}\end{pmatrix},\\
\mathcal{B}^a(\mathbf{w}_{0,E}) &= \frac{1}{4}\Lambda_E\bar{R}^{n_k}\begin{pmatrix}7n_k\mathcal{C}_{k,4}+(2n_k+1)\mathcal{C}_{k,2}+\mathcal{C}_{k,-2}+3(n_k-2)\mathcal{C}_k\\ 7n_k\mathcal{S}_{k,4}+\mathcal{S}_{k,2}-\mathcal{S}_{k,-2}-3n_k\mathcal{S}_k\end{pmatrix}.
\end{aligned}$$

In summary, the total stress is the sum of the stress components due to anisotropy in the elastic constants obtained above, plus the sum for isotropic solids given in [13] multiplied by $\mu = 1 - \frac{H}{2}\frac{1-2\nu}{1-\nu}$, which is reproduced as follows for completeness:

$$\begin{aligned}
\text{(3.23a)}\quad \begin{pmatrix} \sigma_{rr}^{\text{tot,iso}} \\ \sigma_{\phi\phi}^{\text{tot,iso}} \\ \sigma_{r\phi}^{\text{tot,iso}} \end{pmatrix} &= \mu C_1 \begin{pmatrix} 1 - \left(\frac{r}{\bar{R}}\right)^2 \\ 1 - 3\left(\frac{r}{\bar{R}}\right)^2 \\ 0 \end{pmatrix} + \mu\epsilon C_1 \beta_k n_k (n_k - 1) \left(\frac{r}{\bar{R}}\right)^{n_k - 2} \begin{pmatrix} \mathcal{C}_k \\ -\mathcal{C}_k \\ -\mathcal{S}_k \end{pmatrix} \\
&\quad + \mu\epsilon C_1 \beta_k (n_k + 1) \left(\frac{r}{\bar{R}}\right)^{n_k} \begin{pmatrix} (2 - n_k)\mathcal{C}_k \\ (n_k + 2)\mathcal{C}_k \\ n_k \mathcal{S}_k \end{pmatrix}
\end{aligned}$$

and

$$\text{(3.23b)}\quad \sigma_{zz}^{\text{tot,iso}} = 2\mu C_1 \left(1 - 2\left(\frac{r}{\bar{R}}\right)^2\right) + 4\mu\epsilon\beta_k \left(\nu(n_k + 1)C_1 - \frac{1-\nu}{n_k} C_2\right) \left(\frac{r}{\bar{R}}\right)^{n_k} \mathcal{C}_k.$$

3.3. The von Mises and total resolved stresses. A characteristic amount of stress can be assigned to each point with the von Mises stress, which satisfies

$$\text{(3.24)}\qquad 2\sigma_{\text{vm}}^2 = (\sigma_{rr} - \sigma_{\phi\phi})^2 + (\sigma_{rr} - \sigma_{zz})^2 + (\sigma_{\phi\phi} - \sigma_{zz})^2 + 6\sigma_{r\phi}^2$$

for a cubic material, where the σ_{ij} are given by (3.23) and the corrections due to material anisotropy, such as that given by (3.22).

The preferred method of dislocation generation in all III–V semiconductors is through the generation of slip defects, in particular the $\{111\}$, $\langle 1\bar{1}0\rangle$ slip system [1], which consists of four glide planes within which atoms can slip in one of three directions. The resolved stress σ_{rs}, in a particular slip direction $\mathbf{d}$ within the glide plane with normal $\mathbf{n}$, is given by

$$\sigma_{\text{rs}} = \mathbf{d}^{\text{T}} U_{\mathbf{p}}^{\text{T}} Q^{\text{T}} \sigma^{\text{tot}} Q U_{\mathbf{p}} \mathbf{n}.$$

The matrix $U_{\mathbf{p}}$ rotates vectors from the crystallographic frame to the solidification frame. If the stress tensor σ^{tot} is expressed in the (r, ϕ, z) coordinates, Q is the coordinate transformation matrix that takes $(x, y, z) \to (r, \phi, z)$.

Plastic deformation of the crystal occurs if the stress in any of the 12 slip directions exceeds a maximum value known as the critical resolved shear stress, σ_{crss}. To leading order, the actual density of dislocations suffered by the crystal is proportional to the total excess stress at any given point within the crystal. In this sense, an estimation of where dislocations are likely to occur is given by the distribution of the total absolute resolved stress

$$\text{(3.25)}\qquad |\sigma_{\text{rs}}^{\text{tot}}| = \sum_{i=1}^{12} \left| \mathbf{d}_i^T U_{\mathbf{p}}^T Q^T \sigma^{\text{tot}} Q U_{\mathbf{p}} \mathbf{n}_i \right|.$$

4. Numerical results. The physical parameters used for the simulations correspond to InSb grown with the Cz method can be found in [1]. Numerical results are obtained for a conic crystal with a half opening angle of $\varphi_{\text{cone}} = 5°$ so that in nondimensional coordinates $\bar{R}(z) = \bar{R}(Z_0) + \hat{\theta}_{\text{cone}} z$. The initial seed length is $Z_0 = 0.054$ and the radius is $\bar{R}(Z_0) = 1/6$, corresponding to an initial dimensional radius and length of 0.005 m and 0.01 m, respectively. Here we have taken $h_{\text{gs}} = \bar{h}_{\text{gs}} = 4$ so that by using the characteristic radius $\tilde{R} = 0.03$ m and thermal conductivity of

TABLE 1

The maximum von Mises and resolved stress values for the three seed orientations using j correction terms.

	Maximum von Mises stress		
j	$\mathbf{p} = [001]$	$\mathbf{p} = [\bar{1}\bar{1}\bar{1}]$	$\mathbf{p} = [\bar{2}11]$
0	3.32×10^{-3}	7.23×10^{-3}	3.83×10^{-3}
1	2.75×10^{-3}	6.80×10^{-3}	3.89×10^{-3}
12	2.85×10^{-3}	6.89×10^{-3}	4.04×10^{-3}

	Maximum resolved stress		
j	$\mathbf{p} = [001]$	$\mathbf{p} = [\bar{1}\bar{1}\bar{1}]$	$\mathbf{p} = [\bar{2}11]$
0	9.23×10^{-3}	1.34×10^{-2}	8.78×10^{-3}
1	7.66×10^{-3}	1.20×10^{-2}	8.78×10^{-3}
12	8.15×10^{-3}	1.21×10^{-2}	8.84×10^{-3}

4.57 $\mathrm{W\,m^{-1}K^{-1}}$ we find $\mathrm{Bi} = 0.026$. The final radius and length (not including the seed) are 0.03 m and 0.286 m or 1 and 1.537, respectively, in scaled units. This gives a value of $\hat{\theta}_{\text{cone}} = 0.542$.

Θ_0 is the solution of (2.6) in the pseudosteady case ($1/\mathrm{St} = 0$) with $\delta = \gamma = 0$ and $I_{0,1}/I_{2,0} = 1$. Θ_1 is given by (2.7) with $h_F = 0$ so that $F(\Theta) = \beta\Theta = \Theta$. Since the stiffness constants for InSb are $C_{11} = 6.70 \times 10^4$, $C_{12} = 3.65 \times 10^4$, $C_{44} = 3.02 \times 10^4$ MPa, one has $H = 2C_{44} - C_{11} + C_{12} = 2.99 \times 10^4$ MPa and $\omega = 0.329$. In addition, the values of E and ν used in the calculation are represented by (3.2b),

$$E = \frac{(C_{11} + 2C_{12} + H/2)(C_{11} - C_{12} + H/2)}{C_{11} + C_{12} + H/2} = 5.95 \times 10^4 \text{ MPa}, \tag{4.1a}$$

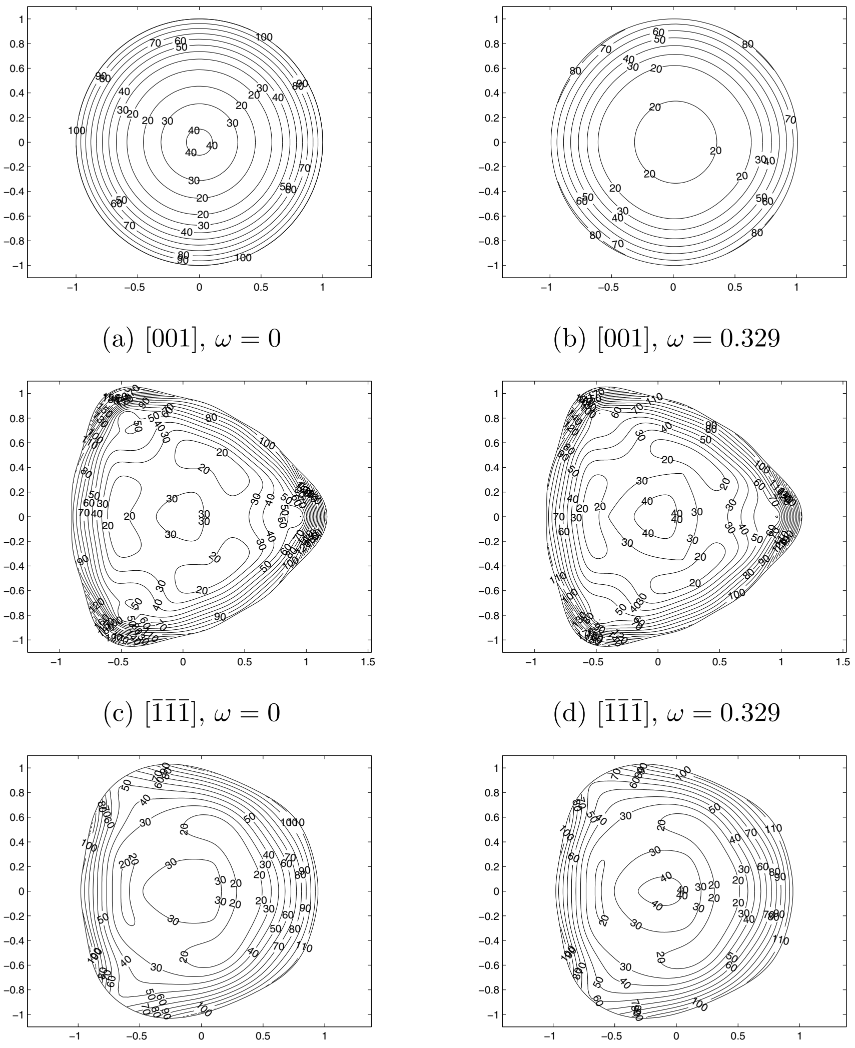
$$\nu = \frac{C_{12}}{C_{11} + C_{12} + H/2} = 0.308. \tag{4.1b}$$

When combined with the parameters above, the dimensional constant for the stress calculations is $\alpha \Delta T E/(1 - \nu) \sim 93.8$ MPa, where $\alpha = 5.5 \times 10^{-6}\mathrm{K}^{-1}$ and $\Delta T = 198.4$ K [1].

We start with the expression for the displacement (3.8), which defines $D^{(1)}$, $D^{(3)}$, $D^{\pm}$, k, and n for the $\mathcal{L}^a(\mathbf{w}_0)$ and $\mathcal{B}^a(\mathbf{w}_0)$ expressions found in sections 3.2.1, 3.2.2, and 3.2.3. The $\mathcal{L}^a(\mathbf{w}_0)$ defines f_r, f_ϕ, k, and $\tilde{n}$ in (3.10), which gives σ^p_{ij}, and $\mathcal{B}^a(\mathbf{w}_0)$ defines g_r, g_ϕ, k, and $\tilde{n}$ in (3.11), which gives σ^h_{ij} with $\tilde{g}_r$ and $\tilde{g}_\phi$ given by g_r, g_ϕ, and σ^p_{ij}.

Figure 1 shows the von Mises stress for the three seeding orientation: [001], $[\bar{1}\bar{1}\bar{1}]$ and $[\bar{2}11]$. To the left of each pair is the isotropic case corresponding to the material in [13], and to the right is the anisotropic case corresponding to one correction term, namely, $\mathbf{w} \sim \mathbf{w}_0 + \mathbf{w}_1$. Reported stress values are given in percentages with 100% corresponding to the outer edge of a cylindrical crystal ($\alpha = 0$) grown in the [001] direction or $|\sigma_{\text{vm}}| = 3.32 \times 10^{-3}$. In the [001] pulling direction the von Mises stress retains its axial symmetry even when anisotropic stiffness coefficients are included. For the $[\bar{1}\bar{1}\bar{1}]$ and $[\bar{2}11]$ seeding orientations the geometric effect dominates the amount of stress. Table 1 lists the maximum value of the von Mises stress for the three orientations using zero (isotropic, $\omega = 0$), one, and twelve correction terms for the total stress. It can be seen that the von Mises stress can either decrease or increase when material anisotropy is considered, depending on seed orientation.

Figure 2 shows the corresponding resolved stress $\sigma^{\text{tot}}_{\text{rs}}$ as given by (3.25), which is relevant to dislocation generation. The computed peak values for the total resolved stress are listed in Table 1 and once again we conclude that the effect of the material



(a) [001], $\omega = 0$

(b) [001], $\omega = 0.329$

(c) $[\bar{1}\bar{1}\bar{1}]$, $\omega = 0$

(d) $[\bar{1}\bar{1}\bar{1}]$, $\omega = 0.329$

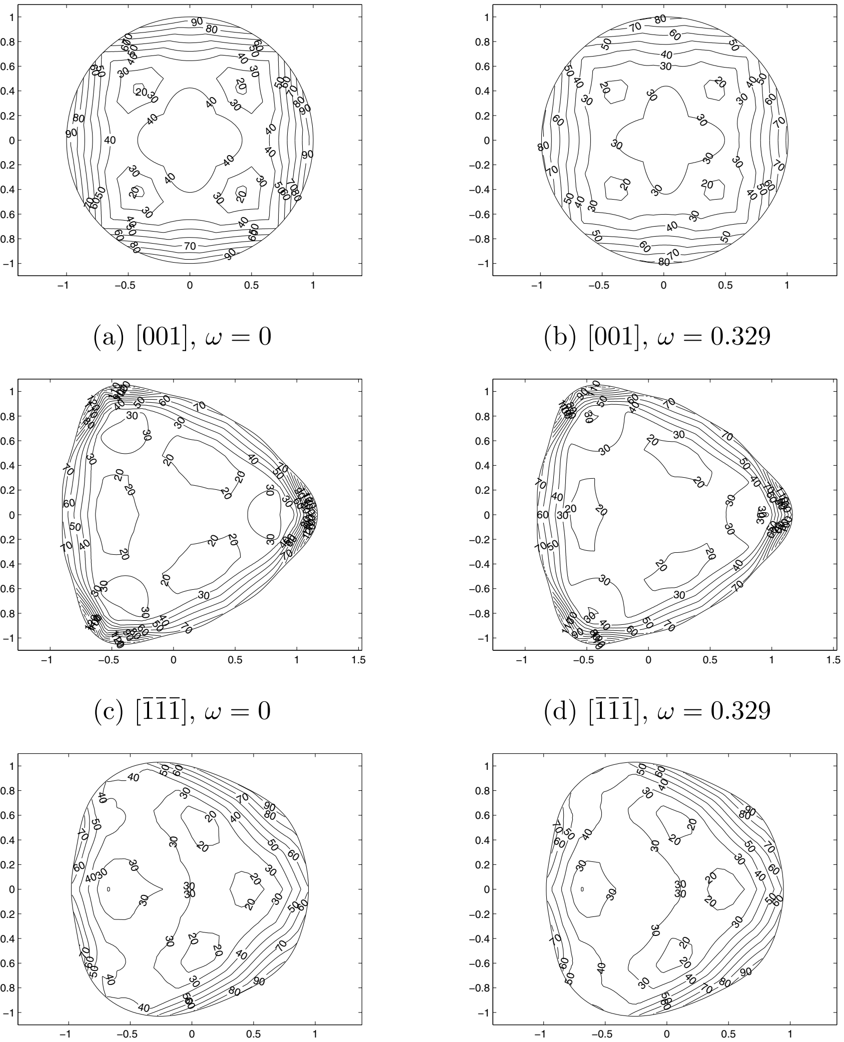
(e) $[\bar{2}11]$, $\omega = 0$

(f) $[\bar{2}11]$, $\omega = 0.329$

FIG. 1. *The von Mises stress computed using* (3.24) *at the indicated orientation, just inside the crystal/melt interface at the end of the growth. All reported stress values are expressed in percentages with* 100% *occurring at the outer edge of a crystal grown in the* [001] *direction, which corresponds to a value of* $|\sigma_{vm}| = 3.32 \times 10^{-3}$(0.311 MPa)*. The* $\omega = 0.329$ *case utilizes one correction term.*

anisotropy is more significant for the [001] orientation since there is no geometric effect in that case. For the other two directions, the geometric effect dominates and the material anisotropy has a limited effect.

5. Conclusion. In this paper we have discussed the effect of material anisotropy on the thermal stress and compared it with that of geometric anisotropy due to facet formation. We have presented a systematic procedure which computes the stress iteratively, using an asymptotic series. We have also shown that the series converges for any anisotropic cubic material. Numerical results are obtained for InSb crystals



(a) [001], $\omega = 0$

(b) [001], $\omega = 0.329$

(c) $[\bar{1}\bar{1}\bar{1}]$, $\omega = 0$

(d) $[\bar{1}\bar{1}\bar{1}]$, $\omega = 0.329$

(e) $[\bar{2}11]$, $\omega = 0$

(f) $[\bar{2}11]$, $\omega = 0.329$

FIG. 2. *The total resolved stress computed using* (3.25) *at the indicated orientation, just inside the crystal/melt interface at the end of the growth. All reported stress values are expressed in percentages with* 100% *occurring at the outer edge of a crystal grown in the* [001] *direction, which corresponds to a value of* $|\sigma_{vm}^{tot}| = 9.23\times10^{-3}(0.866\,\mathrm{MPa})$. *The* $\omega = 0.329$ *case utilizes one correction term.*

grown by the Cz method in three pulling directions. When the seed orientation is in the [001] direction, since no facet forms and no geometric anisotropy is present, the material anisotropy has a visible effect on both the von Mises and the total resolved stresses. For the $[\bar{1}\bar{1}\bar{1}]$ and $[\bar{2}11]$ seeding orientations, however, the material anisotropy has only a limited effect, while the geometric (facet formation) has a much stronger effect. Our results suggest that for faceted crystals, it is much more important to take the geometric effect into account, while neglecting the material anisotropy is justified. Finally, the methodology used in this paper is not limited to the case of Cz crystals.

It can be applied to other elastic problems, such as the ones investigated in [7, 14], by including the effect of an anisotropic material property.

Appendix A. Proof of (3.7). We begin by introducing the weighted direct sum Hilbert space

$$\overset{\circ}{\mathcal{H}} = \left\{ \mathbf{w} \in \left(H^1(\Omega)\right)^3 : \int_\Omega \mathbf{w} \,\mathrm{d}V = \mathbf{0}, \int_\Omega \operatorname{rot}(\mathbf{w}) \,\mathrm{d}V = \mathbf{0}, \|\mathbf{w}\|^2_{\overset{\circ}{\mathcal{H}}} = \sum_{i=1}^{6} \lambda_i \|e_i(\mathbf{w})\|_0^2 \right\}$$

on the bounded domain Ω. $\|\cdot\|_0$ denotes the standard L^2 norm, and the quantity $\mathbf{e}(\mathbf{w})$ consists of elements of the strain tensor associated with the displacement $\mathbf{w}$. The weights λ_i take on the value $C_{11}-C_{12}+H/2$ for $i=1,2,3$ and $C_{44}-H/4$ for $i=4,5,6$ with $H = 2C_{44} - C_{11} + C_{12} \neq 0$ for an anisotropic, cubic material. In Cartesian coordinates, $\mathbf{e}(\mathbf{w}) = (e_{xx}, e_{yy}, e_{zz}, 2e_{yz}, 2e_{xz}, 2e_{xy})^{\mathrm{T}}$, where $e_{ij} = (w_{i,j} + w_{j,i})/2$, the comma denoting partial differentiation.

For the uniqueness of (3.3), (3.5), and (3.6), we assume that the displacement solutions to these equations belong to $\overset{\circ}{\mathcal{H}}$.

Lemma A.1. *For an anisotropic cubic material characterized by the stiffness values* $\{C_{11}, C_{12}, C_{44}\}$*, the quantity*

$$\omega = \frac{|2C_{44} - C_{11} + C_{12}|}{2C_{44} + C_{11} - C_{12}}$$

satisfies $0 < \omega < 1$.

Proof. The eigenvalues of the stiffness matrix $C_{11}+2C_{12}$, $C_{11}-C_{12}$, and C_{44} must be positive, for otherwise the crystal would be unstable [10]. Due to the positivity constraint, we have the strict inequalities

$$-2C_{44} - C_{11} + C_{12} < 2C_{44} - C_{11} + C_{12} < 2C_{44} + C_{11} - C_{12}$$

so that $|2C_{44} - C_{11} + C_{12}| < 2C_{44} + C_{11} - C_{12}$ or $\omega < 1$. The case $\omega = 0$ corresponds to an isotropic crystal. □

The space $\overset{\circ}{\mathcal{H}}$ is the natural choice for an anisotropic cubic crystal. The next lemma states that convergence in $\overset{\circ}{\mathcal{H}}$ is equivalent to convergence in $(H^1)^3$.

Lemma A.2. $\|\cdot\|_{\overset{\circ}{\mathcal{H}}}$ *is equivalent to* $\|\cdot\|_1$ *(the* $(H^1)^3$ *norm) in* $\overset{\circ}{\mathcal{H}}$.

Proof. This is a direct consequence of the Korn inequality [4],

$$\|\mathbf{w}\|_1^2 \le C(\Omega) \left(\sum_{i=1}^{6} \|e_i(\mathbf{w})\|_0^2 \right) \quad \forall \mathbf{w} \in \overset{\circ}{\mathcal{H}},$$

where $C(\Omega)$ is a constant depending only on the domain Ω. □

Next we illustrate that the operator $\mathcal{N}$ is a contraction mapping on $\overset{\circ}{\mathcal{H}}$.

Lemma A.3. *The operator* $\mathcal{N}$ *in* (3.7) *satisfies* $\|\mathcal{N}\|_{\overset{\circ}{\mathcal{H}} \mapsto \overset{\circ}{\mathcal{H}}} \le \omega < 1$.

Proof. For any given $\mathbf{u} \in \overset{\circ}{\mathcal{H}}$, let $\mathbf{w} = \mathcal{N}\mathbf{u}$. Using the boundary condition in the definition of $\mathcal{N}$, we see that $\mathbf{w}$ satisfies

$$\int_\Omega C^0_{ij} e_i(\mathbf{w}) e_j(\mathbf{v}) \,\mathrm{d}V = \int_\Omega C^a_{ij} e_i(\mathbf{u}) e_j(\mathbf{v}) \,\mathrm{d}V \quad \forall \mathbf{v} \in \overset{\circ}{\mathcal{H}},$$

and in particular for $\mathbf{v} = \mathbf{w}$,

$$\int_\Omega C^0_{ij} e_i(\mathbf{w}) e_j(\mathbf{w})\, \mathrm{d}V = \int_\Omega C^a_{ij} e_i(\mathbf{u}) e_j(\mathbf{w})\, \mathrm{d}V. \tag{A.1}$$

Taking only the diagonal terms of the left-hand side of (A.1) yields

$$\int_\Omega C^0_{ij} e_i(\mathbf{w}) e_j(\mathbf{w})\, \mathrm{d}V \geq \int_\Omega \sum_{k=1}^{6} \lambda_k e_k^2(\mathbf{w})\, \mathrm{d}V = \|\mathbf{w}\|^2_{\overset{\circ}{\mathcal{H}}}, \tag{A.2}$$

while noting that C^a is itself diagonal gives

$$\int_\Omega C^a_{ij} e_i(\mathbf{u}) e_j(\mathbf{w})\, \mathrm{d}V \leq \int_\Omega \left(\sum_{k=1}^{3} \left| \frac{H}{2} e_k(\mathbf{u}) e_k(\mathbf{w}) \right| + \sum_{k=4}^{6} \left| \frac{H}{4} e_k(\mathbf{u}) e_k(\mathbf{w}) \right| \right) \mathrm{d}V. \tag{A.3}$$

Using the definitions of ω and H, one has

$$\omega = \frac{|H/2|}{C_{11} - C_{12} + H/2} = \frac{|H/4|}{C_{44} - H/4}$$

so that estimates (A.2) and (A.3) with (A.1) allow us to conclude, with Hölder's inequality, that

$$\|\mathbf{w}\|^2_{\overset{\circ}{\mathcal{H}}} \leq \omega \|\mathbf{w}\|_{\overset{\circ}{\mathcal{H}}} \|\mathbf{u}\|_{\overset{\circ}{\mathcal{H}}}$$

or $\|\mathbf{w}\|_{\overset{\circ}{\mathcal{H}}} \leq \omega \|\mathbf{u}\|_{\overset{\circ}{\mathcal{H}}}$ for any given $\mathbf{u} \in \overset{\circ}{\mathcal{H}}$. Using Lemma A.1, $\|\mathcal{N}\|_{\overset{\circ}{\mathcal{H}} \mapsto \overset{\circ}{\mathcal{H}}} \leq \omega < 1$. □

PROPOSITION A.4. *Let* $\mathbf{s}_n = \mathbf{w}_0 + \mathcal{N}\mathbf{w}_0 + \mathcal{N}^2\mathbf{w}_0 + \cdots + \mathcal{N}^n\mathbf{w}_0$, *with* $\mathbf{s}_0 = \mathbf{w}_0$. *Expression* (3.7) *converges to* $\mathbf{w}$ *in* $\overset{\circ}{\mathcal{H}}$, *and*

$$\|\mathbf{w} - \mathbf{s}_n\|_{\overset{\circ}{\mathcal{H}}} \leq \omega^{n+1} \|\mathbf{w}\|_{\overset{\circ}{\mathcal{H}}}.$$

Proof. Lemma A.3 implies that the right-hand side of (3.7) converges. What remains is to show that $\mathbf{w}$ is in fact the limit. We note that

$$\begin{aligned} \mathcal{L}^0(\mathbf{w} - \mathbf{w}_0) &= \mathcal{L}(\mathbf{w}) + \mathcal{L}^a(\mathbf{w}) - \mathbf{F} = \mathcal{L}^a(\mathbf{w}), && \mathbf{x} \in \Omega, \quad t > 0, \\ \mathcal{B}^0(\mathbf{w} - \mathbf{w}_0) &= \mathcal{B}(\mathbf{w}) + \mathcal{B}^a(\mathbf{w}) - \mathbf{g} = \mathcal{B}^a(\mathbf{w}), && r = R(\phi, z). \end{aligned}$$

By the definition of $\mathcal{N}$, one has $\mathbf{w} - \mathbf{w}_0 = \mathcal{N}\mathbf{w}$ and $\|\mathbf{w} - \mathbf{w}_0\| = \|\mathcal{N}\mathbf{w}\| \leq \omega \|\mathbf{w}\|$. $\mathbf{w} - \mathbf{s}_n = \mathcal{N}(\mathbf{w} - \mathbf{s}_{n-1})$ gives $\|\mathbf{w} - \mathbf{s}_n\| \leq \omega \|\mathbf{w} - \mathbf{s}_{n-1}\|$ with all norms taken in $\overset{\circ}{\mathcal{H}}$. By induction on n

$$\|\mathbf{w} - \mathbf{s}_n\|_{\overset{\circ}{\mathcal{H}}} \leq \omega^{n+1} \|\mathbf{w}\|_{\overset{\circ}{\mathcal{H}}} \quad \forall n \geq 0$$

so that letting $n \to \infty$ and using Lemma A.3 gives the result. □

Appendix B. Detailed form of $C_{\mathbf{cyc}}$. For the $[\bar{1}\bar{1}\bar{1}]$ pulling direction we obtain $C^{a,[\bar{1}\bar{1}\bar{1}]}_{\text{cyc}} = C^{a,[\bar{1}\bar{1}\bar{1}]}_{\text{cyc},0} + C^{a,[\bar{1}\bar{1}\bar{1}]}_{\text{cyc},3}$, where

$$C^{a,[\bar{1}\bar{1}\bar{1}]}_{\text{cyc},0} = \frac{H}{12}\begin{pmatrix} 0 & 2 & 4 & 0 & 0 & 0 \\ 2 & 0 & 4 & 0 & 0 & 0 \\ 4 & 4 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \tag{B.1a}$$

$$C^{a,[\bar{1}\bar{1}\bar{1}]}_{\text{cyc},3} = \frac{\sqrt{2}H}{6}\begin{pmatrix} 0 & 0 & 0 & s_3 & -c_3 & 0 \\ 0 & 0 & 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ s_3 & -s_3 & 0 & 0 & 0 & c_3 \\ -c_3 & c_3 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 0 & c_3 & s_3 & 0 \end{pmatrix}. \tag{B.1b}$$

For the [001] pulling direction, $C^{a,[001]}_{\text{cyc}} = C^{a,[001]}_{\text{cyc},0} + C^{a,[001]}_{\text{cyc},4}$, where

$$C^{a,[001]}_{\text{cyc},0} = \frac{H}{4}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{B.2a}$$

$$C^{a,[001]}_{\text{cyc},4} = \frac{H}{4}\begin{pmatrix} c_4 & -c_4 & 0 & 0 & 0 & -s_4 \\ -c_4 & c_4 & 0 & 0 & 0 & s_4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -s_4 & s_4 & 0 & 0 & 0 & -c_4 \end{pmatrix}. \tag{B.2b}$$

Finally, for the $[\bar{2}11]$ pulling direction, $C^{a,[\bar{2}11]}_{\text{cyc}} = C^{a,[\bar{2}11]}_{\text{cyc},0} + C^{a,[\bar{2}11]}_{\text{cyc},1} + C^{a,[\bar{2}11]}_{\text{cyc},2} + C^{a,[\bar{2}11]}_{\text{cyc},3} + C^{a,[\bar{2}11]}_{\text{cyc},4}$, where

$$C^{a,[\bar{2}11]}_{\text{cyc},0} = \frac{H}{16}\begin{pmatrix} 1 & 3 & 4 & 0 & 0 & 0 \\ 3 & 1 & 4 & 0 & 0 & 0 \\ 4 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \tag{B.3a}$$

$$C^{a,[\bar{2}11]}_{\text{cyc},1} = \frac{\sqrt{2}H}{24}\begin{pmatrix} 0 & 0 & 0 & -s_1 & 3c_1 & 0 \\ 0 & 0 & 0 & -3s_1 & c_1 & 0 \\ 0 & 0 & 0 & 4s_1 & -4c_1 & 0 \\ -s_1 & -3s_1 & 4s_1 & 0 & 0 & c_1 \\ 3c_1 & c_1 & -4c_1 & 0 & 0 & -s_1 \\ 0 & 0 & 0 & c_1 & -s_1 & 0 \end{pmatrix}, \tag{B.3b}$$

$$C^{a,[\bar{2}11]}_{\mathrm{cyc},2} = \frac{H}{24}\begin{pmatrix} -2c_2 & 0 & 2c_2 & 0 & 0 & s_2 \\ 0 & 2c_2 & -2c_2 & 0 & 0 & s_2 \\ 2c_2 & -2c_2 & 0 & 0 & 0 & -2s_2 \\ 0 & 0 & 0 & -2c_2 & -2s_2 & 0 \\ 0 & 0 & 0 & -2s_2 & 2c_2 & 0 \\ s_2 & s_2 & -2s_2 & 0 & 0 & 0 \end{pmatrix}, \tag{B.3c}$$

$$C^{a,[\bar{2}11]}_{\mathrm{cyc},3} = \frac{\sqrt{2}H}{8}\begin{pmatrix} 0 & 0 & 0 & s_3 & -c_3 & 0 \\ 0 & 0 & 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ s_3 & -s_3 & 0 & 0 & 0 & c_3 \\ -c_3 & c_3 & 0 & 0 & 0 & s_3 \\ 0 & 0 & 0 & c_3 & s_3 & 0 \end{pmatrix}, \tag{B.3d}$$

$$C^{a,[\bar{2}11]}_{\mathrm{cyc},4} = \frac{7H}{48}\begin{pmatrix} -c_4 & c_4 & 0 & 0 & 0 & s_4 \\ c_4 & -c_4 & 0 & 0 & 0 & -s_4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ s_4 & -s_4 & 0 & 0 & 0 & c_4 \end{pmatrix}. \tag{B.3e}$$

REFERENCES

[1] C. S. Bohun, I. Frigaard, H. Huang, and S. Liang, *A semianalytical thermal stress model for the Czochralski growth of type* III–V *compounds*, SIAM J. Appl. Math., 66 (2006), pp. 1533–1562.

[2] W. A. Brantley, *Calculated elastic constraints for stress problems associated with semiconductor devices*, J. Appl. Phys., 44 (1973), pp. 534–535.

[3] T. Chen, H. Wu, and C. Weng, *The effect of interface shape on anisotropic thermal stress of bulk single crystal during Czochralski growth*, J. Crystal Growth, 173 (1997), pp. 367–379.

[4] C. O. Horgan, *Korn's inequalities and their applications in continuum mechanics*, SIAM Rev., 37 (1995), pp. 491–511.

[5] D. T. J. Hurle, *Handbook of Crystal Growth*, Vols. 1 & 2, North-Holland, Amsterdam, 1994.

[6] J. C. Lambropoulos, *The isotropic assumption during the Czochralski growth of single semiconductors crystals*, J. Crystal Growth, 84 (1987), pp. 349–358.

[7] Y.-H. Lin, *A higher order asymptotic analysis for orthotropic plates in stress edge conditions*, J. Elasticity, 77 (2004), pp. 25–55.

[8] W. F. H. Micklethwaite, *The bulk growth of InSb and related ternary alloys*, in Bulk Crystal Growth of Electronic, Optical and Optoelectronic Materials, P. Capper, ed., John Wiley & Sons, Hoboken, NJ, 2005, Chapter 5.

[9] N. Miyazaki, *Development of a thermal stress analysis system for anisotropic single crystal growth*, J. Crystal Growth, 236 (2002), pp. 455–465.

[10] J. F. Nye, *Physical Properties of Crystals*, Oxford University Press, Oxford, UK, 2001.

[11] S. Vigdergauz and D. Givoli, *Thermoelastic stresses in a crystal with weak anisotropy*, J. Crystal Growth, 198/199 (1999), pp. 125–128.

[12] S. Vigdergauz and D. Givoli, *Thermoelastic stresses in a cylinder or disk with cubic anisotropy*, Internat. J. Solids and Structures, 36 (1999), pp. 2109–2125.

[13] J. Wu, C. S. Bohun, and H. Huang, *A thermal elastic model for constrained crystal growth with facets*, submitted.

[14] G. W. Young and J. A. Heminger, *A mathematical model of the edge-defined film-fed growth process*, J. Engrg. Math., 38 (2000), pp. 371–390.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 305–329

A CONTINUUM MODEL FOR A CHUTE FLOW OF GRAINS*

A. S. ELLIS† AND F. T. SMITH†

Abstract. This paper is motivated by a problem from the food-sorting industry and is concerned with the development of a continuum model for a chute flow of grains, based on analogies with the Lighthill–Whitham model of traffic flow. A fundamental relationship between the density and the flux is proposed which is multivalued due to the fact that grains can move leftwards and rightwards across the chute. The subsequent implications are discussed in detail. Results are presented first by solving with a method of characteristics and second by solving analytically and numerically after including a viscous dissipation term to smooth out discontinuities. The fundamental diagram is revisited in view of the viscous dissipation term and a substantial modification at the cusps is found necessary to allow physically sensible solutions to emerge. If the viscous dissipation is small, then in general the viscous effects also are small. Exceptions, however, are shocks, where the viscosity has a smoothing effect, as is well known, and two new features/observations: crossovers, which allow two-way solutions, and steady states, which can exist only in the viscous case and appear over a long time scale.

Key words. wave equation, shocks and related discontinuities, granular flows

AMS subject classifications. 35L05, 74J40, 76T25

DOI. 10.1137/080713057

1. Introduction.

The aim of this paper is to define and explore a mathematical model for the nearly two-dimensional (2D), gravity-driven, rapid flow of a monolayer of grains down an inclined chute. This is directly motivated by a problem from the food-sorting industry, in particular from a company that manufactures machines for the sorting of food, Sortex Ltd.

In the particular food-sorting process developed by Sortex, grains fall from a hopper and are subsequently moved along by a vibrator tray. At the end of the tray the grains fall onto an inclined chute, down which they are accelerated due to gravity. They quickly form an apparent 2D monolayer upon the chute. Shortly after the grains have fallen from the bottom of the chute they pass an optical system that can detect defective grains. A grain is considered to be defective if it is, for example, of the wrong size, shape, or color. Foreign bodies, such as small shards of glass, can also be detected. If a defective grain or foreign body is detected, a powerful jet of air is fired from at least one ejector in an array of ejectors, and the grain is knocked into a reject bucket by the force of the impact. A schematic diagram of the process is shown in Figure 1.1. Studies of the ejector and jet properties are in theses by Westwood [16] and Wilson [18].

The chute is approximately 30cm wide and a meter in length, and its angle of inclination from the vertical is 30°. The grains exit the chute with a vertical velocity of the order of 4–5ms^{-1}. A typical grain of rice has a width of 1–3mm and a length of 5–7mm. The mass has a magnitude of roughly 10^{-5}kg.

Particular difficulties arise as the grains fall off the chute, since they are not uniformly distributed but typically clustered and inhomogeneous. As a consequence, the air jet can, and usually does, remove other grains of rice surrounding the target

*Received by the editors January 10, 2008; accepted for publication (in revised form) July 7, 2008; published electronically October 29, 2008. This work was partially supported by Sortex, Ltd.

http://www.siam.org/journals/siap/69-2/71305.html

†Department of Mathematics, UCL, Gower Street, London, WC1E 6BT, UK (aellis@math.ucl.ac.uk, frank@math.ucl.ac.uk).

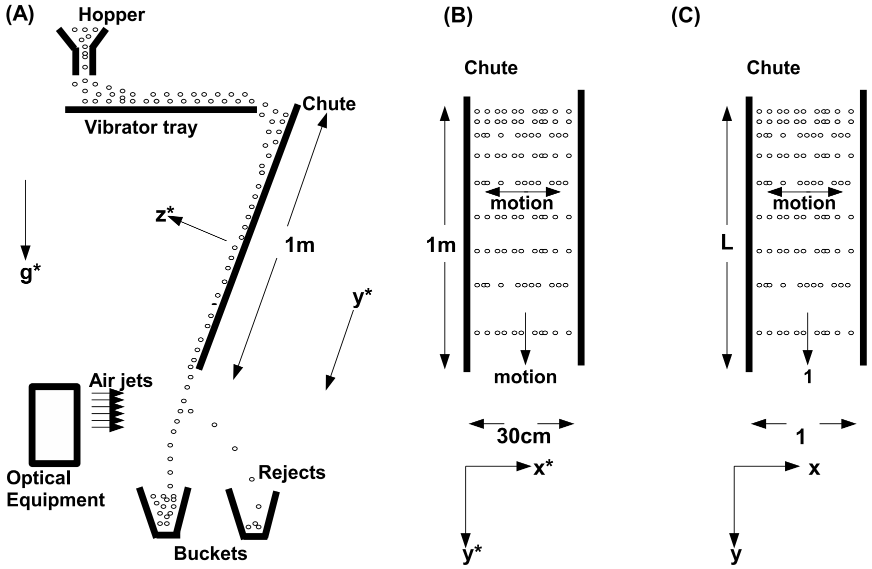


Fig. 1.1. *A schematic diagram of the food-sorting process showing* (A) *the side view of the sorting machine;* (B) *a dimensional front view of the chute depicting the grain motion; and* (C) *a nondimensional front view of the chute. Cartesian coordinates x^*, y^*, z^* are shown and g^* is the dimensional acceleration due to gravity. The diagram is not to scale.*

grain which may not themselves be defective. This is a source of inefficiency in the food-sorting process; the grains in the reject bucket sometimes must be sorted through again to reduce waste. Currently, the optical system can be configured to yield an increased sorting performance, but if a high level of sorting is required a chute with channels must be used; these align the grains with the ejectors and improve the uniformity of the product feed. There is, however, a concomitant reduction in the mass of grains that can be sorted in a given time (known as the "throughput"). The main goal is to increase uniformity of "product feed" as much as possible while maintaining a high throughput of grains. In the current work we therefore try to find a suitable mathematical model for a chute flow of grains so that the above issue regarding ejection can be better treated.

As there is no general theory applicable to chute flows of grains, we necessarily develop a theory from modeling the process from first principles. Other attempts have been made, notably the approach of Jenkins and Savage [9], who describe a *kinetic theory* of the particle motion, whereas after consideration we choose to develop a *continuum description* of the chute flow. It may seem unusual to model an inherently discrete process with a continuum model, but there is a strong and successful history of doing so in other particulate flow problems. For example, continuum models have been used to model traffic and pedestrian flows [7], [8], [10], [11], [12] and many aspects of granular flows; see, for example, the review by Rajchenbach [14] and the works of Aranson and Tsimring [1], [2] and Grossman, Zhou, and Ben-Naim [6] and the references therein, to name but a few. The particulate nature of the problem can be essentially overlooked by addressing larger macroscale behavior and assuming that quantities such as density are continuous. It is then possible to write hydrodynamic-like equations that govern the overall particulate motion. For example, Hughes [7] used a continuum theory of pedestrian flows to suggest a method of safely placing barriers to prevent tragic events as pilgrims cross the bridge of Jamarat in Mecca, and Lighthill and Whitham [11] used a continuum theory of kinematic waves to describe traffic flow. Their results explain well-observed phenomena such as traffic jams at traffic lights and density waves in moving vehicular flow. In this paper we extend the Lighthill–Whitham theory of traffic flow to produce a description of a chute flow.

The structure of the paper is as follows. Basic modeling and simple analytic solutions are dealt with first in sections 2 and 3, including descriptions of separating grains and clashing grains. Next in section 4 more general solutions are sought numerically, and we shall see that this necessitates the introduction of a viscous dissipation term into the governing equations, and that these solutions compare well to those found from the inviscid theory. In section 5 it is demonstrated that in principle two-way solutions can exist for the viscous model in which the grains move both left and right across the chute. Steady states are discussed in section 6 before concluding remarks are made in section 7.

2. Continuum modeling. The continuum model below is based on an argument for extending the Lighthill–Whitham model of traffic flow to cover aspects of the chute flow. Strengths and weaknesses of the argument are described. In particular, the continuum model we propose requires the introduction of a fundamental relation between the nondimensional flux q and the nondimensional density ρ, and the underlying physical mechanism requires discussion. The variables are nondimensionalized as follows: $x^* = a^*x$, $y^* = L^*y$, $v^* = V^*v$, $u^* = U^*u$, $t^* = (a^*/U^*)t$, $\rho^* = (M^*/a^*)\rho$, and $q^* = (M^*U^*/a^*)q$. Here variables with an asterisk are dimensional and variables without an asterisk are nondimensional. The parameters are defined so that a^* is the distance across the chute; L^* is the distance down the chute; V^* is a typical vertical velocity given by $\sqrt{g^*h^*}$, say, where h^* is distance fallen down the chute and g^* is the acceleration due to gravity; finally U^* is a typical velocity across the chute. We follow through the implications of the proposed flux-density relation in detail and an appraisal is given in the conclusion. We cannot deny that the model omits many factors, and as such is incomplete, as would be any first model. In a sense, the approach is an empirical one, and we shall examine whether the model can describe some special situations seen on chutes in reality, although the original physical arguments per se may remain open to question.

We make the assumption that the density of rice grains, ρ, forms a continuum and that the grains move along a horizontal line (the x axis) with a flow rate q. The lines are considered to fall down the chute with increasing time t, their vertical displacement being given by $y^* = -g^*{t^*}^2$ (the grains are assumed to have a vertical speed $v^* = 0$ at the top of the chute where $y^* = 0$). Although the assumption of grain motion occurring only in horizontal lines appears simplistic, because in reality vertical interactions could be expected to be an important mechanism of the flow, as noted below, computations with this assumption achieve physically reasonable results as qualitatively they show the development of clusters and voids which are seen in experiments [5]. Moreover, the one-dimensional (1D) model itself is found to contain some rich and complex behavior suggesting that to try to start immediately with a 2D model might be premature.

In addition to the a posteriori motivation for modeling grain motion along a horizontal line, mentioned above, there exists a priori justification for doing so which is as follows. We shall see shortly that the spatially 2D extension, which includes vertical grain motion, reduces to the 1D model of present interest *at large distances* *down the chute* (see (2.4) below). Primarily we are interested in obtaining the density profile *at the bottom of the chute*, where the ejection issue occurs. Given that the 2D model reduces to the 1D model in the lower portion of the chute, then it seems appropriate to use the 1D equation, at least as a first approximation. In fact, the interest and novelty in this work lie in the perhaps surprising result that the horizontal grain motion is seen to dominate the flow behavior, rather than the vertical motion of grains down the chute. The reason for the existence of horizontal grain motion is described later, but briefly it is due to the action of the vibrator tray.

A continuum model for this discrete process is believed justifiable since in reality a state can evolve on the chute where the grains form into large coalesced masses, each one moving as if it is one body with a particular velocity distribution and each one containing a large number of grains. The density of these masses may be different in each case and as there are large numbers of grains the density could take any value. When the flow is in this state there are sudden jumps in density between each cluster and there can also be voids. We shall see that this can be modeled by the shocks and expansion fans in the continuum model and thus the model provides qualitative agreement with experiments, which do indeed indicate the formation of clusters and voids.

By conservation of mass the continuity equation is

$$\rho_t + q_x = 0. \tag{2.1}$$

Taking q to depend only on ρ keeps the wave problem simple as a first approximation and may also be justified on physical grounds, and if $q_\rho \equiv \frac{dx}{dt}$, then the total derivative $\frac{d\rho}{dt} = 0$. Defining the wavespeed $q_\rho \equiv c\,(\rho)$, the density is constant along straight characteristics given by

$$x = c\,(\rho)\,t + x_0, \tag{2.2}$$

where x_0 is a constant of integration representing an initial position for x. The resultant equation for the flow is

$$\rho_t + c\,(\rho)\,\rho_x = 0. \tag{2.3}$$

Therefore an initial density distribution $\rho = f\,(x)$ at $t = 0$ determines in principle the density evolution with time via the characteristics.

To determine $c\,(\rho)$ we argue that the flux q is related to the density ρ by $q = Q\,(\rho)$ say and then $c\,(\rho) = Q'\,(\rho)$, where the prime denotes differentiation with respect to the argument. (We shall discuss shortly the validity of choosing a particular $q = Q(\rho)$ relation.) We must also initially specify q along each characteristic. The values (ρ, q) in the initial condition thus determine a unique value of c which in turn determines the gradient of the characteristic. The flux and the density are then constant along the characteristic, i.e., both q and ρ are propagated along the characteristics. The initial conditions in the current section are piecewise-constant as these provide a basic starting point for the analysis.

An important point concerning improvement and extension is to make the model spatially 2D, the 2D equation being

$$\rho_t + c_1(\rho)\rho_x + c_2(\rho)\rho_y = 0, \tag{2.4}$$

where y is the vertical distance down the chute. Seeking a solution independent of y as $y \to \infty$ leads back to the 1D model of (2.3), however. Thus the 1D model of

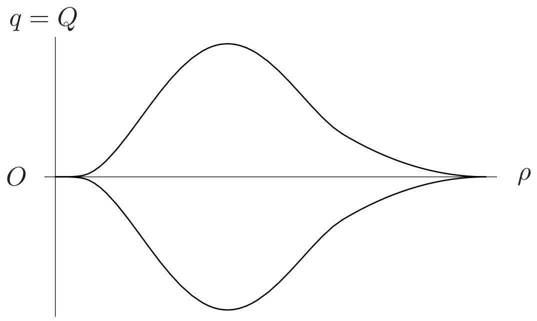


Fig. 2.1. *Sketch of the fundamental diagram, including the cusps at $q = \rho = 0$ and at $q = 0$, $\rho = \rho_M$ and the inflexion points relatively nearby.*

present interest evolves directly from the 2D case at large times or sufficiently large distances down the chute. This justifies a priori the assumption of horizontal grain motion.

We propose below a particular $Q(\rho)$ relation for the chute flow (see Figure 2.1); the validity of the proposition is discussed shortly. Note that, borrowing from traffic flow theories, the $Q(\rho)$ curve is known as the fundamental diagram or fundamental curve. If $\rho = 0$, there can be no flow, so then $q = 0$ trivially. Further, it is argued that there is no flow for a maximum value ρ_M of the density, corresponding to a "jamming" of grains across the chute where each grain is touching the neighboring grain or wall and hence, within some interval of at least x, there is no room for any grain to move relatively across the chute. In between q takes a single maximum at some value of the density ρ_F. However, the rice can travel in both directions (unlike the traffic flow case) and hence q can also take negative values (of the same magnitude by virtue of symmetry) for each value of ρ. The reason for this bidirectional grain motion is as follows. Before the grains fall down the chute, they are transported along a vibrator tray. As they fall off the vibrator tray, its motion induces horizontal velocity fluctuations. Hence, as they travel down the chute, grains can collide, which induces further horizontal velocity fluctuations, or they can separate. Thus there are two branches of the fundamental diagram, one which describes leftward-moving grains and the other rightward-moving grains, and so the flux-density relation is necessarily double-valued.

Finally here, the $Q\,(\rho)$ curve on each branch is expected to pass through inflexion points relatively near the cusps, which lie at the zero-q end of the branches. The two branches meet at cusps so that the wavespeed remains finite and smooth as the solution passes through the endpoints and switches branches; they also allow mass-conserving shock-fan structures, as we shall see later, which enable physically acceptable descriptions of clashing and separating regions to develop. In more detail, the cusps must behave locally as

$$q \propto \pm\rho^{\frac{3}{2}} \text{ near } \rho = 0 \tag{2.5a}$$

or

$$q \propto \pm(\rho_M - \rho)^{\frac{3}{2}} \text{ near } \rho = \rho_M. \tag{2.5b}$$

To see this consider putting $\rho = f(\mu)$, where $\mu = \frac{x}{t}$ is a similarity variable. The governing equation is then satisfied provided that

$$(c - \mu)f' = 0; \tag{2.6}$$

so $\mu = c$ is allowed, which produces expansion fans. Next, at the high-density endpoint (the reasoning applies equally to the zero-density endpoint) consider having locally

$$q = \pm\beta(\rho_M - \rho)^n \tag{2.7}$$

with the unknown power $n > 0$ and the constant β being nonzero. Via the definition of c we obtain from (2.7) the local behavior

$$\rho = \rho_M - \left(\frac{|\mu|}{\beta n}\right)^{\frac{1}{n-1}}. \tag{2.8}$$

Consequently the relation $\frac{1}{n-1} = M$ must hold with M an even integer in order that the wavespeed varies smoothly as the density passes through its maximum. Hence $n = 1 + \frac{1}{M}$. In particular, the value

$$n = \frac{3}{2} \tag{2.9}$$

corresponds to $M = 2$ and the density $\rho = \rho_M - (\frac{|\mu|}{\beta n})^2$, which would be expected to be the most general case. Similar reasoning for a cusp also applies at the low-density end. The fundamental curve's upper branch must therefore be concave upwards at its endpoints. It follows also that for there to be a maximum q in between, *inflexion points* must be produced between the maximum and the endpoints. The only alternative to having cusps, while retaining a smooth wavespeed as the solution switches branches, would be to have an infinite wavespeed at the endpoints, which is physically unacceptable. Flux functions with inflexion points also arise in traffic flow theory; see, for example, Lebacque and Khoshyaran [10] or Morgan [12].

Concerning strengths and weaknesses of the present proposed fundamental diagram, in pedestrian and traffic flows there are obvious physical reasons why people or cars slow down with increased density (overcrowding, driver nervousness, and so on), whereas it is difficult to provide a comprehensive argument as to why, for example, a densely packed region of grains may move more slowly than a sparser region as in the cases here. This difficulty could be countered by arguing that, when the densities are low, grains may have small horizontal velocities because there is a smaller probability of velocities being induced impulsively by a collision. At large densities, however, collisions are likely to be very numerous and thus grain speeds would reduce as energy would be lost repeatedly at each collision. At moderate densities the probability of a collision lies somewhere in between these two extremes and so horizontal velocities may be induced impulsively by a collision, but there may not be so many repeated collisions as to cause a continual reduction in the velocity component via restitution, so that the magnitude of the horizontal velocities may be larger. In the extreme case of a blockage the grains would come to a complete stop and may become packed at the jammed density. Thus a situation arises in which the density influences the flux, or vice versa, and the view that $q = Q(\rho)$ appears to be justified (at least as a first approximation). Further, owing to these simple arguments, a shape of such a fundamental curve similar to the one proposed above seems to be suggested. Alternatively, we could argue that such a choice of fundamental diagram is appropriate for certain physical situations, such as with colliding or separating grains on a chute. It may be of significance here that at a collision the grains can be considered to instantaneously change velocity at the point of touching, and so the flux of the grains is zero when the density is maximal, exactly as in the proposed fundamental diagram. Other situations, such as with a highly dense region moving rapidly on an otherwise empty chute, should be covered by another fundamental diagram, but such a situation may be unlikely to develop in practice because high densities seem more likely to arise when grain speeds are slow. It is also worth mentioning that as clusters and voids are the key feature of the computational results [5] and as collisions and separations are believed to be the crucial mechanism behind the formation of clusters and voids, then this aspect of the flow may be the most pertinent part to consider in an initial model. Some of the above criticisms may also hold for the theory when it is applied to traffic flow: for example, a densely packed region of cars on an otherwise empty highway will not in reality have $q = 0$; they may accelerate away and diffuse.

Given the above setting, we continue with the present 1D formulation and, in brief, examine the outcome.

3. Inviscid solutions. Interesting aspects arise in the model because $Q(\rho)$ is smoothly varying on each branch and so the characteristics generally intersect or diverge within a finite time if ρ and q vary on each characteristic. Intersections are a significant feature since the density is implied as multivalued. Such an apparent contradiction is resolved by the formation of a “shock” (see Whitham [17]) which travels with velocity

$$U = \frac{q_2 - q_1}{\rho_2 - \rho_1}, \tag{3.1}$$

which is the gradient of the chord between (ρ_1, q_1) and (ρ_2, q_2) on the fundamental curve. Diverging characteristics potentially create an area devoid of information about the density but lead to an “expansion fan.” The aim in what follows is to employ the shock wave and expansion fan structures as mechanisms to obtain inhomogeneous density distributions upon the chute and provide some further explanation of clusters and voids when grains are colliding or separating.

A number of simple analytical solutions are illustrated here. We shall see that it is possible to build increasingly complex solutions to the continuum model. In theory, it is possible to determine any solution analytically by examining the characteristics, together with the shocks and fans. The approach is much the same as that of Morgan [12], the difference here being the multivalued fundamental diagram.

Let us first define ρ_F as the value of the density for which q is maximum; ρ_{IR} as the value of the density at the right-hand inflexion point; and ρ_{IL} as the value of the density at the left-hand inflexion point.

3.1. Example solution one. We start simply with a classical example [4] in which a shock must occur. Consider two adjacent regions of constant density, one with density ρ_1 and the other with density ρ_2, and allow $\rho_1 < \rho_2 < \rho_F$. The corresponding values of the flux are q_1 and q_2, respectively, with $0 < q_1 < q_2 < q_F$. It is possible to choose this arrangement such that $c_1 > c_2$; see Figure 3.1 for clarity.

Furthermore, if the region of density ρ_1 is assumed to lie to the left of the region of density ρ_2, as depicted in Figure 3.2(a), then the characteristics are seen to intersect in the $x - t$ plane. As has already been stated, the density at first glance would be multivalued at such an intersection since the density is a different constant along each intersecting characteristic. This would be physically unacceptable.

The resolution is well known, namely, to replace the intersecting points with a shock, i.e., a sudden jump in the density. In this way, we see that the correct $x - t$

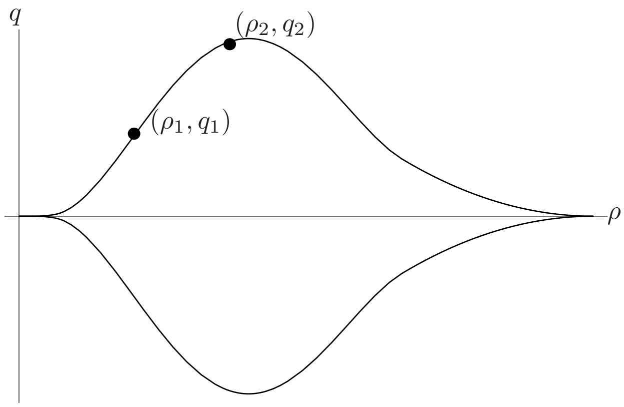
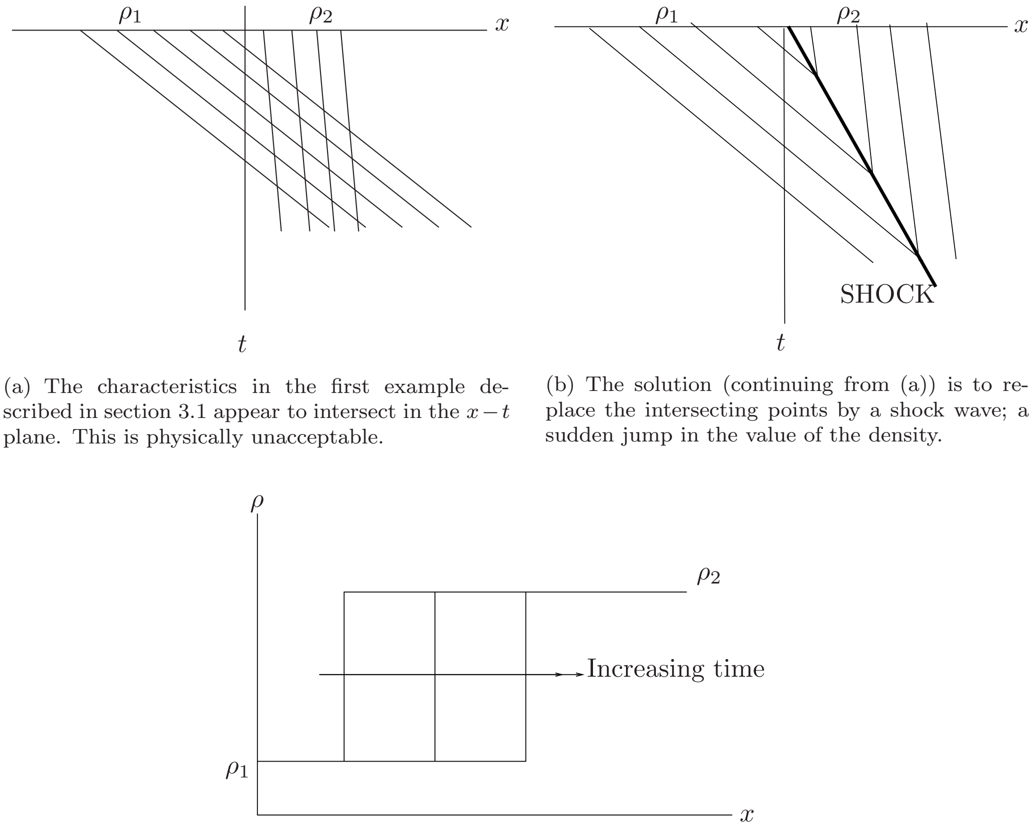


FIG. 3.1. *Illustration highlighting the values of* (ρ_1, q_1), (ρ_2, q_2) *for the case producing a shock outlined in section* 3.1. *The respective slopes of the tangents at* (ρ_1, q_1) *and* (ρ_2, q_2) *are* c_1 *and* c_2.



(a) The characteristics in the first example described in section 3.1 appear to intersect in the $x - t$ plane. This is physically unacceptable.

(b) The solution (continuing from (a)) is to replace the intersecting points by a shock wave; a sudden jump in the value of the density.

(c) The corresponding density profile is shown above as it evolves over time. The two regions translate rightwards, separated by a shock.

FIG. 3.2. *Shock formation.*

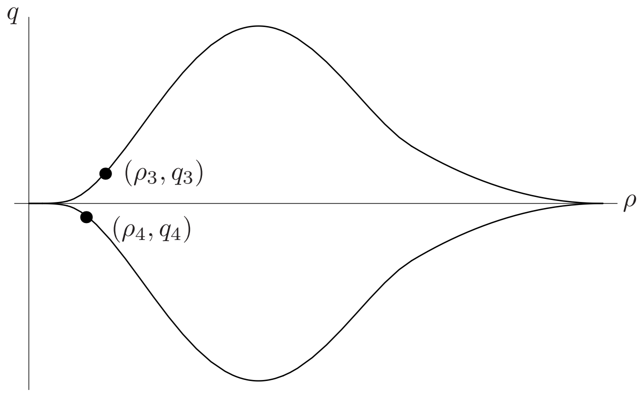


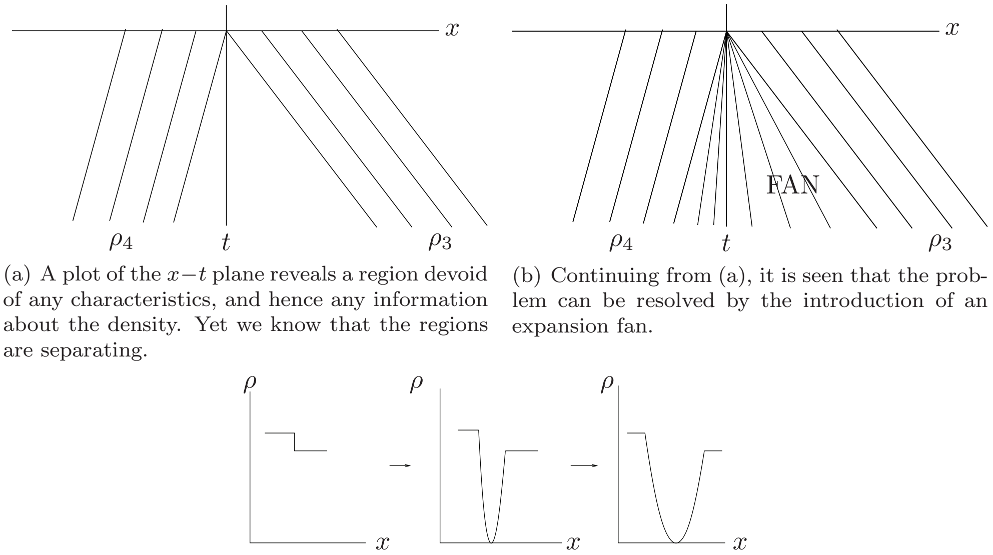
FIG. 3.3. *Illustration highlighting the values of* (ρ_3, q_3), (ρ_4, q_4) *for the case of the "fan" outlined in section* 3.2.

diagram for this case is as in Figure 3.2(b), in which two regions of constant density are separated by a shock.

The corresponding evolution of the density profile is as in Figure 3.2(c). Both regions are moving rightward (as $q > 0$) and there is a translating shock between the two regions.

3.2. Example solution two. Similarly, a scenario which involves an expansion fan is as follows. Again consider two adjacent regions of constant density, one with density ρ_3 and the other with density ρ_4, where the corresponding values of the flux are q_3 and q_4, respectively. However, now suppose that $q_3 > 0$, $q_4 < 0$, and $\rho_4 < \rho_3 < \rho_{IL}$. Thus $c_3 > 0$, $c_4 < 0$, and $|c_3| > |c_4|$; see Figure 3.3. (If $\rho_{IL} < \rho_3 < \rho_4 < \rho_F$, then the solution is a little more complex, as we shall see in a later example.) The region with density ρ_3 is assumed to lie to the right of the region with density ρ_4, as in Figure 3.4(a). Such an arrangement corresponds to the two regions moving apart since $q > 0$ in the right-hand region and $q < 0$ in the left-hand region, and in the $x - t$ diagram there is seen to be a region devoid of characteristics. Consequently, there appears to be no information about the density evolution here, yet we know that the regions are separating. The problem can be resolved by the introduction of an expansion fan.

An expansion fan is a region of characteristics which all start from the same point, but their gradient continuously changes from the value of the gradient of the characteristic in the right-hand region to the gradient of the characteristic in the left-hand region. Hence the void region is now replaced by a "fan" of characteristics whose gradients decrease monotonically. As the gradient varies through this fan, so must the density. See Figure 3.4(b). The continual change in the gradient corresponds to moving from the point (ρ_3, q_3) on the upper branch of the $q(\rho)$ curve to the point (ρ_4, q_4) via the cusp at $(0, 0)$. Notice that the characteristic at the center of the fan has zero slope and thus the point of zero density is stationary. Therefore the fan in this case corresponds to a gradual decrease in the density and a reduction in flux to a stationary central point with zero density. This is followed by a gradual increase in the density accompanied by an increase in magnitude of the flux, which is now negative. The expansion fan has allowed the density to *switch* branches. The physical interpretation of this is indeed a separation of the two regions. Figure 3.4(c) illustrates the evolution of the corresponding density profile to highlight the physics.



(a) A plot of the $x-t$ plane reveals a region devoid of any characteristics, and hence any information about the density. Yet we know that the regions are separating.

(b) Continuing from (a), it is seen that the problem can be resolved by the introduction of an expansion fan.

(c) The evolution of the density profile corresponding to the preceding figure shows that the expansion fan captures the physical process of separation.

Fig. 3.4. *The expansion fan solution of the example in section* 3.2.

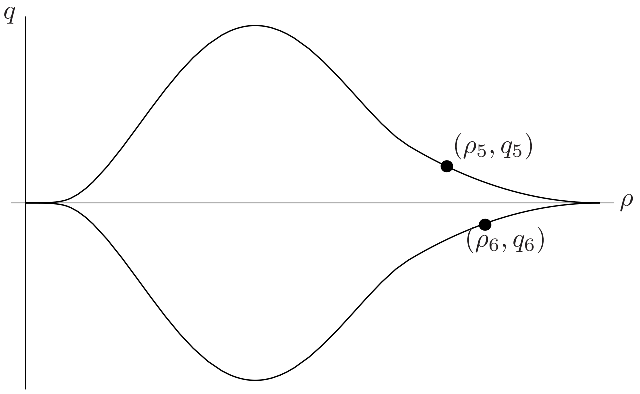


Fig. 3.5. *Illustration highlighting the values of* (ρ_5, q_5), (ρ_6, q_6) *for the example of a fan in a colliding region.*

3.3. Example solution three. This again is a classical problem [4] in which expansion fans arise, this time describing regions of colliding grains. Consider two regions of constant density with $\rho_6 > \rho_5 > \rho_{IR}$ and $q_5 > 0$ and $q_6 < 0$. Now $c_5 < 0$ and $c_6 > 0$ since ρ_5 and ρ_6 lie toward the large density end of the fundamental diagram, as in Figure 3.5. Assuming that the ρ_5 region lies to the left of the ρ_6 region will result in an $x - t$ plot of the characteristics that is qualitatively similar to that in the example above in Figure 3.4. Such an arrangement now corresponds to clashing of grains. Again there will be a region devoid of characteristics where an expansion fan can be introduced. However, in this example, the monotonically decreasing gradient of the characteristics in the fan corresponds to moving along the

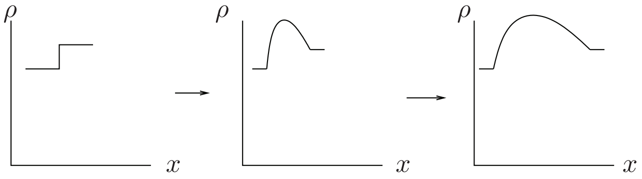


FIG. 3.6. *The corresponding evolution of the density profile for Figure* 3.5 *illustrates the expansion fan capturing the physical process of collision.*

fundamental diagram from (ρ_6, q_6) to (ρ_5, q_5) via the cusp located at $(\rho_M, 0)$. Observe that the characteristic at the center of the fan has zero slope and thus the point of maximal density is stationary.

Thus, moving through the fan from ρ_6 to ρ_5 allows the density to switch branches from the lower branch to the upper branch while passing through a point of maximum density. The fan describes a "hump" of large density at the location where we know that grains are colliding. Figure 3.6 shows the evolution of the density profile to highlight this point.

The expansion fan structures can be described analytically [17]. The characteristics satisfy (2.2) and each characteristic in the fan crosses the x axis at the same point, hence x_0 is the same constant for each one. Therefore we can rearrange (2.2) to find the gradient of each characteristic in the fan as

$$c(\rho) = \frac{x - x_0}{t}. \tag{3.2}$$

Therefore the complete solution for the wavespeed is

$$c = \begin{cases} c_5, & c_5 < \frac{x-x_0}{t}, \\ \frac{x-x_0}{t}, & c_6 < \frac{x-x_0}{t} < c_5, \\ c_6, & \frac{x-x_0}{t} < c_6. \end{cases} \tag{3.3}$$

3.4. Example solution four. A more complex case is illustrated here. Consider two regions of constant density: the right-hand region with $\rho_F < \rho_7 < \rho_{IR}$ and $q_7 < 0$ and the left-hand region with $\rho_F < \rho_8 < \rho_{IR}$ and $q_8 > 0$; see Figure 3.7. Thus $c_7 > 0$ and $c_8 < 0$. If the characteristics are plotted in the $x - t$ plane, there will again be a region devoid of characteristics which we intuitively expect to describe a clashing region.

One simply might expect the solution of this problem to be an expansion fan between (ρ_7, q_7) and (ρ_8, q_8), but it is actually a little more involved, as follows. Since ρ_7 and ρ_8 are to the left of the inflexion point ρ_{IR} the characteristics in an expansion fan would not monotonically decrease from c_7 to c_8. Hence an expansion fan cannot be immediately plotted.

Instead, consideration indicates that there must be a shock from ρ_7 to ρ_{T-} and a shock from ρ_8 to ρ_{T+}, where ρ_{T-} is the point where a chord drawn from (ρ_7, q_7) is tangent to the fundamental curve. Similarly, ρ_{T+} is the point where a chord drawn from (ρ_8, q_8) is tangent to the fundamental curve. This is similar to the approach of Morgan [12]. Both ρ_{T-}, $\rho_{T+} > \rho_{IR}$, clearly. See Figure 3.7. An expansion fan which passes through the cusp at $(\rho_M, 0)$ can now be drawn between ρ_{T-} and ρ_{T+}. Such a "shock-fan-shock" structure will still conserve mass.

Figure 3.8(a) shows the plot of the characteristics in the $x - t$ diagram for this situation, and Figure 3.8(b) depicts the time evolution of the density profile. There

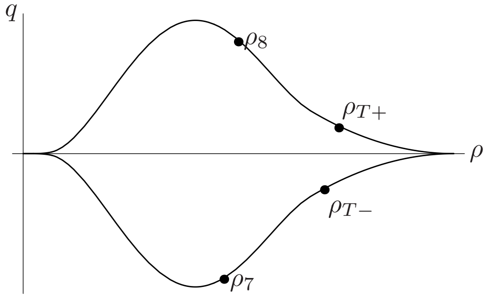
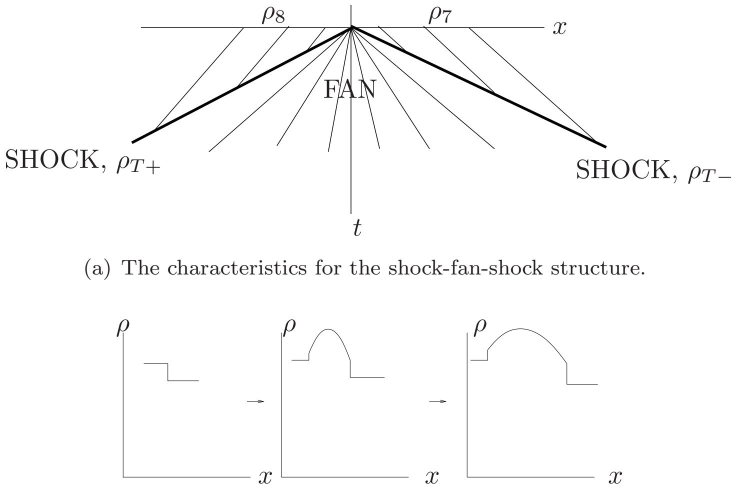


FIG. 3.7. *The values of* (ρ_7, q_7), (ρ_8, q_8), (ρ_{T-}, q_{T-}), (ρ_{T+}, q_{T+}) *for a case including a mix of shocks and fan(s) in a colliding region.*



(a) The characteristics for the shock-fan-shock structure.

(b) The corresponding evolution of the density profile is shown for (a).

FIG. 3.8. *The shock-fan-shock structure of the solution in the example in section* 3.4.

is a region of constant density moving leftward, then a shock to a fan where there is a high density region, then a shock down to a region of constant density moving rightward. Thus the characteristics appear to give a solution where a high-density cluster develops, as might be anticipated for a clashing region.

3.5. Example solution five. Shock-fan-shock structures can arise in other situations. Consider $\rho_9 = \rho_7$, $\rho_{10} = \rho_8$ given as in the above case, but now with ρ_{10} lying in the right-hand region and ρ_9 lying in the left-hand region, so that the regions are separating. At first sight, it seems that the characteristics are intersecting and so the solution ought to be a shock. However, if a line were drawn through the intersecting points, the gradient would not be equal to the gradient of the chord between (ρ_9, q_9) and (ρ_{10}, q_{10}); the chord would not have the required speed U (hence such a shock would not conserve mass and moreover, by violating the Rankine–Hugoniot condition, it would not be a weak solution to the conservation law).

The problem is avoided by the introduction of two shocks: one from ρ_9 to ρ_{T2-}, and the other from ρ_{10} to ρ_{T2+}. The point ρ_{T2-} is the place on the fundamental curve where the gradient c_{T2-} is tangent to a chord drawn from ρ_9 such that $\rho_{T2-} < \rho_{IL}$.

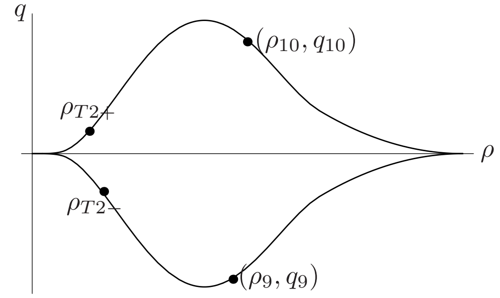
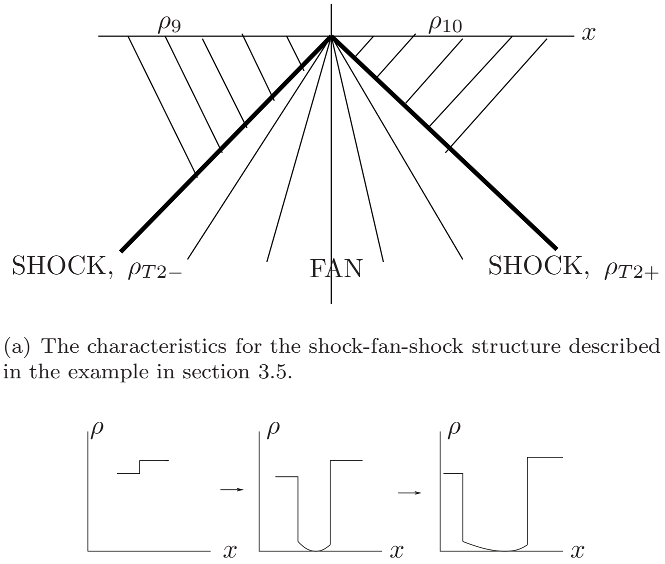


FIG. 3.9. *The values of* (ρ_9, q_9), (ρ_{10}, q_{10}), (ρ_{T2-}, q_{T2-}), (ρ_{T2+}, q_{T2+}) *for a case including a mix of shocks and fan in a separating region.*



(a) The characteristics for the shock-fan-shock structure described in the example in section 3.5.

(b) A sketch of the time evolution of the density profile for the example of a shock-fan-shock structure when grains are separating.

FIG. 3.10. *The shock-fan-shock structure of the solution in the example in section* 3.5.

Similarly, the point ρ_{T2+} is the place on the fundamental curve where the gradient c_{T2+} is tangent to a chord drawn from ρ_9 such that $\rho_{T2+} < \rho_{IL}$. See Figure 3.9 for details. An expansion fan can now be drawn between ρ_{T2-} and ρ_{T2+} that switches branches through the cusp at zero density. Thus there is a region of constant density moving leftward adjacent to a leftward-translating shock down to small densities. Then there is a fan through zero density that is next to a rightward-moving shock that jumps to large densities; see Figure 3.10. Hence the characteristics here appear to give a solution in which a void develops, as might be anticipated for a separating region.

3.6. Example solution six. It is useful to consider an example with three discontinuous density regions to show the complexity of possible solutions. Thus consider three regions of density ρ_{11}, ρ_{12}, and ρ_{13} as given in Figure 3.11. These are chosen such that $\rho_{IL} < \rho_{11} < \rho_F$ with corresponding $q_{11} > 0$; $\rho_{12} > \rho_{IR}$ with

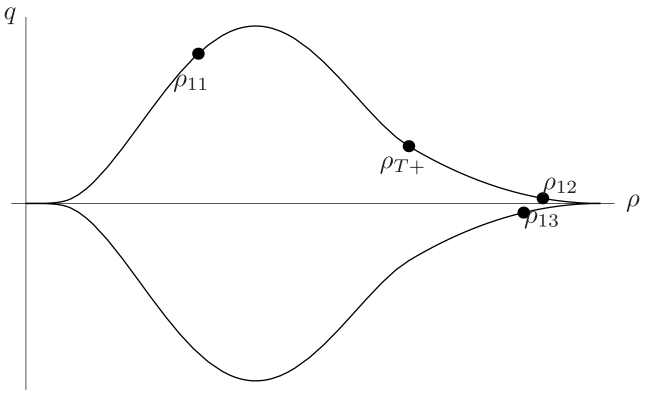


FIG. 3.11. *The important values of ρ_{11}, ρ_{12}, etc., are shown on the fundamental diagram for the example in section* 3.6*, which has three regions of different density.*

corresponding $q_{12} > 0$; and $\rho_{13} > \rho_{IR}$ with corresponding $q_{13} < 0$. We also draw attention to the point ρ_{T+}, which is the point where a chord drawn from (ρ_{11}, q_{11}) is tangent to the fundamental curve on the upper branch. Observe that $\rho_{IL} < \rho_{T+} < \rho_{IR}$ and the chord is the shock s_1.

These three regions are assumed to lie in the $x - t$ plane such that the region of density ρ_{13} lies to the right of the region of density ρ_{12}, which in turn lies to the right of the region of density ρ_{11}. Thus we observe that there is an area of colliding grains between ρ_{13} and ρ_{12}, and consequently an expansion fan is required here. Further consideration shows that a fan is also required at the bottom of the ρ_{12} region to a density with value ρ_{T+}. The characteristic with density ρ_{T+} coincides with the shock s_1 where there is a jump down to density ρ_{11}. Consequently the $x - t$ plane is as shown in Figure 3.12(a), and a schematic representing the evolution of the density profile is shown in Figure 3.12(b).

These examples illustrate how shocks and fans can be used to model clusters and voids. We wish to highlight here that we have observed that there might be cases, for example, where waves of negative speed and positive velocity occur and where waves of positive speed but negative velocity occur. The physical interpretation of this, as in the example in section 3.3, is when two regions collide, causing grains to “pile up” and form a blockage. This is analogous, say, to a traffic jam in a unidirectional traffic flow model. The transition between positive and negative velocities is either grains colliding and hence reversing velocities or grains which are separating due to the initial velocities induced by the vibrator tray. More complex solutions can be constructed in a similar fashion for any number of colliding or separating regions, and interactions with walls may be included. We also remark here that in general a steady state behavior does not set in at, say, large times for the inviscid model, except for ρ, q constant.

A further point for discussion is that the model above could instead be written as

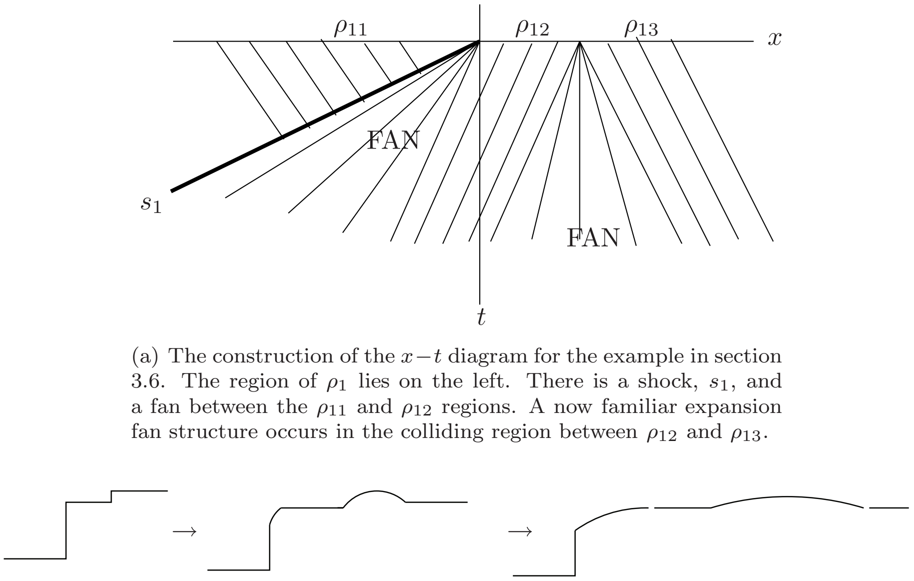
$$\rho_t + Q(q, \rho)_x = 0 \tag{3.4a}$$

and

$$q_t = 0, \tag{3.4b}$$

where

$$Q(q, \rho) = \mathrm{sgn}(\rho) f(\rho), \tag{3.4c}$$



(a) The construction of the $x - t$ diagram for the example in section 3.6. The region of ρ_1 lies on the left. There is a shock, s_1, and a fan between the ρ_{11} and ρ_{12} regions. A now familiar expansion fan structure occurs in the colliding region between ρ_{12} and ρ_{13}.

(b) A sketch of the the evolution of the density profile corresponding to (a). Density is on the vertical axis, and the position across the chute is along the horizontal axis. A shock exists between ρ_{11}, on the left of the chute, and the small fan to the region of density ρ_{12}. A "hump" of high density exists in the colliding region between ρ_{12} and ρ_{13}. As time increases the two fans spread out. The shock on the left persists for all time.

FIG. 3.12. *The solution from the example in section* 3.6*, which involves three regions of different density.*

where f is a classical flux function. The six examples above could be examined in the realm of the above system, which yields a 2×2 conservation law whose Riemann problems can be analyzed. This could form the scope of future work, and we are grateful to a referee for making this observation.

4. Viscous effects. Only examples that have discontinuous initial conditions have so far been investigated. Obviously we would wish to find the long-term density distributions for an arbitrary set of initial conditions, and in particular we would like to solve for problems with continuous initial density distributions. However, owing to the shock wave and expansion fan structures, it is difficult to do this in general, both analytically and computationally.

We now attempt to find such generalized solutions, but in this section we consider only those which *remain always on one branch* of the fundamental diagram, that is, only those situations in which *grain movement is unidirectional* and especially those which are not near the chute walls. (Solutions which require a branch switch, in which grains move to and fro, are dealt with in the following section.)

One standard way to compute general solutions to equations of the above form is to add an artificial viscous dissipation term $\nu\rho_{xx}$, so all discontinuities can be "smoothed out" in principle, as the equation is now parabolic. Indeed, this is the conventional method used in traffic flow problems [13], [17]. Whitham [17] discusses at length the validity of such an approach and shows that in the limit of the viscosity ν tending to zero the solutions do in fact asymptote toward the familiar shock and fan structures seen for the inviscid equation.

A strong physical argument for including viscous dissipation in our case is the inclusion of air effects. As grains approach each other in collisions, effects of air cushioning can reduce the importance of impacts in the model for certain flow rates [5], [15] and thus make the density distribution more homogeneous. Another line of argument is that including a viscous dissipation term is comparable to including the next term in a Taylor expansion of the flux, in the sense that now the flux $q = Q(\rho) - \nu\rho_x$, where ν is a small positive constant; see Whitham [17]. Equation (2.3) is therefore modified to

$$\rho_t + c\rho_x = \nu\rho_{xx}. \tag{4.1}$$

This admits some general solutions for the density in a one-way flow, which we investigate below. Equation (4.1) is of central importance here and is referred to frequently hereafter as the "continuum equation."

The continuum equation (4.1) is solved numerically by a finite difference scheme. In the examples below, $\nu = 0.0001$. Upwind or downwind differencing is used for the spatial first derivative, depending on the sign of $c(\rho)$ at each cell in the numerical grid. Although this method of solution is standard and well known, we wish to briefly examine two problems in the new context of the chute flow.

An appropriate flux-density relation must first be specified for the continuum model. We take

$$q = \begin{cases} \pm\rho^4, & 0 \le \rho < 1, \\ \pm\left(c_1\rho^4 + c_2\rho^3 + c_3\rho^2 + c_4\rho + c_5\right) & 1 \le \rho < 10, \\ \pm\left(\rho_M - \rho\right)^2, & 10 \le \rho < \rho_M, \end{cases} \tag{4.2}$$

where for now $\rho_M = 15$; $c_1, c_2, \dots, c_5$ are constants chosen to ensure that the function matches smoothly, and ± obtains the upper or lower branch, respectively. Equation (4.2) captures the main features required of the fundamental diagram that were elucidated earlier. As we consider only unidirectional grain movement we choose the positive branch without loss of generality.

We illustrate two solutions here. Recall that in section 3 solutions were for discrete, discontinuous input, as these formed a basic starting point for the analysis. We attempt to approximate such initial conditions in the code, although of course it is not particularly desirable to start the computation with discontinuous input.

For the first case here we find a solution which mimics the translating shock type of solution.

The initial condition used for this case is

$$\rho(x,0) = \begin{cases} 2 + e^{-25}, & x \le 4, \\ e^{-25(x-5)^2} + 2, & 4 < x \le 5, \\ 3, & x > 5, \end{cases} \tag{4.3}$$

and the boundary conditions are

$$\rho(0,t) = 2 + e^{-25}, \tag{4.4a}$$

$$\rho(15,t) = 3. \tag{4.4b}$$

Hence there are two regions of constant density, both with $q > 0$ (i.e., the density is fixed on the upper branch), and there is a smooth transition between the two

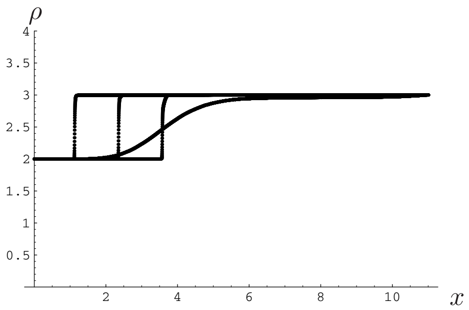


FIG. 4.1. *The initial condition together with solutions at $t = 0.5,\ 1,$ and 1.5. The initial condition is the smoothest thick black line. As time increases, the "jump" between the two near-constant regions steepens considerably and translates leftward, mimicking a translating shock. The solution is shown at times $t = 0.5,\ 1,$ and 1.5.*

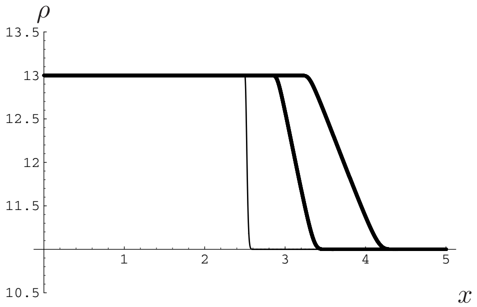


FIG. 4.2. *The initial condition and two solutions, one at $t = 0.1$ and the other at $t = 0.2$. As time increases, the "jump" between the two near-constant regions flattens considerably and spreads rightward across the chute, mirroring well an expansion fan.*

regions; this approximates the discontinuous input leading to a translating shock type of solution in the $\nu = 0$ "inviscid" case.

Figure 4.1 shows the initial condition together with solutions obtained at times $t = 0.5,\ 1,$ and 1.5. As time increases, the "jump" between the two near-constant regions steepens considerably and translates leftward, mimicking well a translating shock.

For the second case here an expansion fan type of solution is replicated. The initial condition is

$$\rho(x,0) = \begin{cases} 13, & x \le 2.5, \\ 2e^{-1000(x-2.5)^2} + 11, & 2.5 < x \le 3, \\ 11 + 2e^{-250}, & x > 3. \end{cases} \tag{4.5}$$

The boundary conditions are

$$\rho(0,t) = 13 + 2e^{-250}, \tag{4.6a}$$

$$\rho(5,t) = 11. \tag{4.6b}$$

Hence initially there are two regions of constant density, and a smooth, yet steep, transition between the two regions.

Figure 4.2 shows the initial condition and the solution obtained at times $t = 0.1$ and 0.2. As time increases, the "jump" between the two near-constant regions flattens considerably and spreads rightward across the chute, resembling an expansion fan, as hoped. Other solutions can similarly be found.

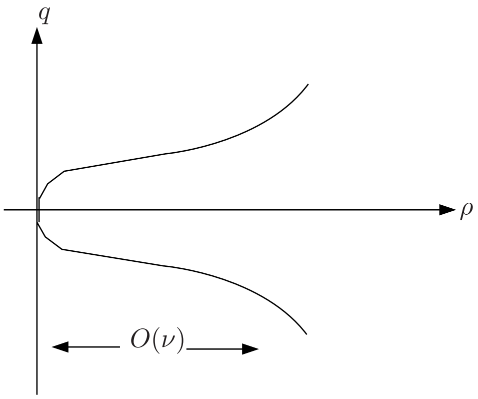


FIG. 5.1. *A sketch of the new fundamental diagram in the neighborhood of $\rho = q = 0$.*

5. Viscous two-way flows. In this section, we demonstrate that in principle valid solutions can be obtained for the continuum equation in which grains can move to and fro across the chute; i.e., there exist valid solutions for "viscous" two-way flows. Such solutions, however, require a modification to the fundamental diagram at the endpoints. In particular, we stated in section 2 that the two branches of the fundamental diagram must meet at cusps at the endpoints. Such geometry is required if the wavespeed is to change smoothly during the branch switch *in the inviscid model*, but in the viscous parabolic governing equation such discontinuities are found to be incompatible with the viscous dissipation term. Consequently, the local inviscid $q \sim \rho^{\frac{3}{2}}$ relation should be altered for the viscous case.

On changing the fundamental diagram locally near the endpoints, we show below that the local behavior near the origin (or maximum) must in general be

$$q \propto \rho^{\frac{1}{2}} \text{ near } \rho = 0 \tag{5.1a}$$

or

$$q \propto (\rho_M - \rho)^{\frac{1}{2}} \text{ near } \rho = \rho_M, \tag{5.1b}$$

and then we examine if a valid solution exists. The suggested behavior above is required by the following reasoning. As the solution passes through a minimum, as it must do in separating regions, $\rho \propto x^2$ is clearly the most general case. This, in answer to a referee's query, is because we expect ρ to be small and positive as the solution passes through a minimum. As x is also small (since ρ is expected to be zero in the center of a separating region) the higher order terms x^4, x^6, etc., give only an asymptotically small correction, in general. Balancing terms in (4.1) thus reveals immediately that $q \sim x$, i.e., $q \propto \rho^{\frac{1}{2}}$ as just above. A similar relation applies near a maximum. The consequences of this new law in the vicinity of $\rho = 0$ are considered next, where we aim to find a similarity solution valid at *small time*. Figure 5.1 shows a sketch of the new fundamental diagram in the neighborhood of $\rho = q = 0$. Note that the size of the viscous region is $O(\nu)$.

Locally $q \sim 2c_0\rho^{\frac{1}{2}}$, say, with c_0 a constant, so $c \sim c_0\rho^{-\frac{1}{2}}$ and substitution into (4.1) yields the ordering

$$\frac{\rho}{t} \pm c_0\rho^{-\frac{1}{2}}\frac{\rho}{x} \sim \frac{\rho}{x^2}.$$

Balancing terms in ρ we see that, in terms of orders, $x \sim t^{\frac{1}{2}}$ and hence $\rho \sim t$. This is acceptable since the density is small in a separating region. Therefore the similarity variable $\eta = \frac{x}{t^{\frac{1}{2}}}$ and the form $\rho = tf\left(\eta\right)$ hold locally.

It is also clear that application of this local argument is actually quite wide-ranging because of the parameters c_0, ν. Their inclusion gives the representative size of x as $(\nu t)^{\frac{1}{2}}$ and hence the size of ρ as $\frac{c_0^2 t}{\nu}$, implying formally that the small-density law continues to apply provided that $t \ll \frac{\nu}{c_0^2}$. The time range could therefore be small or large, depending on the parameters.

The solutions for the current revised model thus have regions which are governed by the new viscous behavior when $\rho \approx 0$ or $\rho \approx \rho_M$, and regions governed by the inviscid behavior away from $\rho \approx 0$ or $\rho \approx \rho_M$, where the fundamental diagram is unchanged. As $\nu \to 0$ the viscous regions effectively vanish and the fully inviscid solutions are regained. In this way, the solutions from the inviscid model and the current viscous model are correlated.

Substituting into the continuum equation (4.1) and choosing $+c_0$ with c_0 positive when the flux q is greater than zero, where locally $f' > 0$, and choosing $-c_0$ with c_0 positive when the flux q is less than zero, where $f' < 0$ locally, we obtain the nonlinear ordinary differential equation

$$\nu f'' - \left(\pm\frac{c_0}{f^{\frac{1}{2}}} - \frac{\eta}{2}\right) f' - f = 0 \tag{5.2}$$

as the nominal small-time equation near the density extremum. It is shown in the coming analysis that f is necessarily zero at a minimum, a physically sensible result for the density in the center of a separating region. Thus (5.2) is to be solved subject to $f(k) = f'(k) = 0$, where $\eta = k$ is the location of the minimum. Also f is expected to be a smooth function for all η and to grow proportionally to η^2 at large $|\eta|$.

As a check, approximating the behavior away from the extremum, in the core, we put $c \sim \pm c_0 \rho^{-\frac{1}{2}}$ and expand

$$\rho = \rho_0\left(x\right) + t\rho_1\left(x\right) + t^2\rho_2\left(x\right) + \cdots . \tag{5.3}$$

Now if we assume that, for some positive constant λ, $\rho_0 \approx \lambda x^2$, which is the most general form for a minimum (separating grains) local to the origin, then on substitution into (4.1) we obtain

$$\rho = \lambda x^2 + 2t\lambda^{\frac{1}{2}}\left(\lambda^{\frac{1}{2}}\nu \mp c_0\right) - t^2\frac{c_0^2 \mp c_0\nu\lambda^{\frac{1}{2}}}{x^2} + O(t^3). \tag{5.4}$$

Hence if $x^2 \sim t$, then the three leading terms become $O\left(t\right)$ and the series is no longer asymptotic. This reinforces the earlier similarity variable $\eta = xt^{-\frac{1}{2}}$. Moreover, if $\lambda = c_0^2\nu^{-2}$, then the $O\left(t\right)$ and the $O\left(t^2\right)$ terms are zero and so the expansion may still be valid. Therefore a simple crossover between branches is possible with $\lambda = c_0^2\nu^{-2}$.

Indeed, turning to (5.2) we observe that

$$f = \Gamma\eta^2 \tag{5.5}$$

is an exact solution if $\Gamma = \left(\frac{c_0}{\nu}\right)^2$. All other solutions fall into one of two categories. One is where the minimum of f is zero, in which case a series solution through the minimum is required so the numerical scheme does not blow up. The second category is where the minimum has f nonzero: the series in this case is not regular. The latter might be dismissed by a physical argument (the density must be zero at the center of a separating region) but an analysis is presented for completeness. The series are helpful in the subsequent numerical study.

First we put $f = F^2$; so (5.2) becomes

$$\nu F F'' + F' \left(\nu F' \mp c_0 + \frac{1}{2} (k + s) F \right) - \frac{1}{2} F^2 = 0, \tag{5.6}$$

where we have defined $s \equiv \eta - k$ and $f' = 0$ at $\eta = k$, a prime denoting differentiation with respect to s.

We mention here that having k nonzero allows the minimum, which cannot then be at the origin, to move with speed $\dot{x} = \frac{k}{2} t^{-\frac{1}{2}}$. Thus the solution has a fixed minimum point at $x = 0$ only if $k = 0$ (we shall see later that this corresponds to the exact solution $f = \Gamma \eta^2$). We are free to choose k in the local problem; it is actually determined by the global solution across the whole chute. Since $t \ll 1$ slower movements correspond to $k \to 0$, in effect, and faster movements can be roughly approximated by $|k|$ becoming large.

We must find a series solution through the minimum point, in order that a Runge–Kutta scheme can start away from the minimum point, where (5.2) has a singularity. We therefore write an expansion for F near the minimum point:

$$F = F_0 + s F_1 + s^2 F_2 + \cdots . \tag{5.7}$$

Since $f' = 0$ at $\eta = k$ then $2FF' = 0$ at $s = 0$, implying that $F_0 F_1 = 0$. In the first instance we choose $F_1 = 0$ and hence

$$F = F_0 + s^2 F_2 + s^3 F_3 + O(s^4). \tag{5.8}$$

Substituting (5.8) into (5.6) reveals that at leading order

$$F_0 = 0 \qquad \text{or} \qquad F_2 = \frac{F_0}{4\nu}. \tag{5.9}$$

So this leads to either $f = 0$ at the minimum or to a series for the case when $f = F_0$ at the minimum. If the former case is chosen, then we can re-expand

$$F = s F_1 + s^2 F_2 + s^3 F_3 + s^4 F_4 + O(s^5) \tag{5.10}$$

about $s = 0$, since the condition $F_0 F_1 = 0$ is automatically satisfied (we shall return shortly to the case $F_0 \neq 0$, $F_1 = 0$, $F_2 = \frac{F_0}{4\nu}$). Expansion (5.10) ultimately results in

$$F = \frac{c_0}{\nu} \left(s - s^2 \frac{k}{8\nu} + s^3 \frac{k^2}{96\nu^2} - s^4 \frac{1}{256} \left(\frac{k^3}{6\nu^3} - \frac{k}{\nu^2} \right) + \cdots \right) \tag{5.11}$$

to the right of the minimum, and

$$F = -\frac{c_0}{\nu} \left(s + s^2 \frac{k}{8\nu} + s^3 \frac{k^2}{96\nu^2} + s^4 \frac{1}{256} \left(\frac{k^3}{6\nu^3} - \frac{k}{\nu^2} \right) + \cdots \right) \tag{5.12}$$

to the left of the minimum. In fact, any number of terms in the series can be deduced.

We can now use the series in (5.11) to march from the minimum to the right to some positive value $s = a$, say (i.e., $\eta = k + a$) to obtain f and f' there. Similarly we can use the series in (5.12) to find the values of f and its first derivative at some $s = -b$, $b > 0$, which is to the left of the minimum ($\eta = k - b$). Hence we have the starting conditions for two Runge–Kutta schemes: one starting at $\eta = k + a$ and solving (5.2) shooting forward to some large positive η, and the other starting at $\eta = k - b$ and solving (5.2) shooting backward to some large negative η. In the following solutions we have also normalized both c_0, ν to one.

Before examining these solutions, however, we return to the option where the minimum $f \neq 0$, i.e., $F_0 \neq 0$ and $F_1 = 0$, giving the alternative result that $F_2 = \frac{F_0}{4\nu}$ at leading order (5.9). In this case the expansion to the right of the minimum is

$$F = F_0\left(1 + \frac{s^2}{4\nu} + s^3\left(-\frac{k}{24\nu^2} + \frac{c_0}{12\nu^2 F_0}\right)\right) \tag{5.13}$$

and the expansion to the left of the minimum is

$$F = F_0\left(1 + \frac{s^2}{4\nu} + s^3\left(-\frac{k}{24\nu^2} - \frac{c_0}{12\nu^2 F_0}\right)\right). \tag{5.14}$$

The central point here is that the third term in (5.13) differs from the third term in (5.14) in the sign of c_0. The series about the minimum is therefore not regular if $F_0 \neq 0$, which is unacceptable as we are seeking a smooth solution. An inner-inner region would be required if the series solution were nonregular, within which more knowledge of the local physics would be required, possibly concerning "jump conditions," for example. Hence it turns out that the series with this choice is not regular and furthermore it results in the unphysical condition that $f \neq 0$ at the center of a separating region, and this option is therefore ultimately dismissed. Altogether, therefore, the minimum must occur with $f = 0$, and so series (5.11) and (5.12) can be used in conjunction with a Runge–Kutta method to find solutions of (5.2). These correspond to the density being zero at the center of a separating region, agreeing with physical intuition.

The first solution we find is for the values $a = b = 0.1$ and $k = 1$, corresponding to the minimum being located at $(\eta, f) = (1, 0)$. A series expansion is used to find the solution between $\eta = [0.9, 1.1]$, and then two Runge–Kutta schemes are used to shoot forwards or backwards from the endpoints of the series. Figure 5.2 shows that the density increases relatively rapidly to large values to the left of the minimum yet increases to a smaller value to the right of the minimum: the density distribution is asymmetric either side of the minimum. Four terms have been used in the series expansion in this case. Although this may seem a surprisingly small number, when the number of terms in the series is changed to check the numerical accuracy the solution remains virtually the same.

The numerical accuracy of the above solution was further checked. The step-length was made shorter or longer to check the grid-dependence of the solution. Also the number of terms in the series expansion was changed, and the interval of η in which the series is applied, to make sure the solution is not dependent on either. Finally, the length of the series can be changed to further ensure the result has no dependence on this as well. The solution is found to be robust to the above changes, and we can conclude that the numerical scheme is indeed sufficiently accurate.

Other solutions can be found where the minimum is located at different points. These solutions have quantitatively different behavior from the one found above. For example, the solution shown in Figure 5.3 has the minimum placed at $(\eta, f) = (0, 0)$. Again, a four-term series expansion is used to find the solution through the minimum,

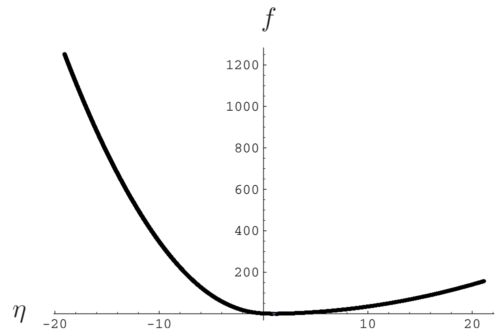


FIG. 5.2. *A solution to* (5.2) *where the minimum occurs at* $(\eta, f) = (1, 0)$. *A series solution has been used to enable passage through the minimum. The step size in this example is* $h = 0.001$. *The grains separate into regions of differing density.*

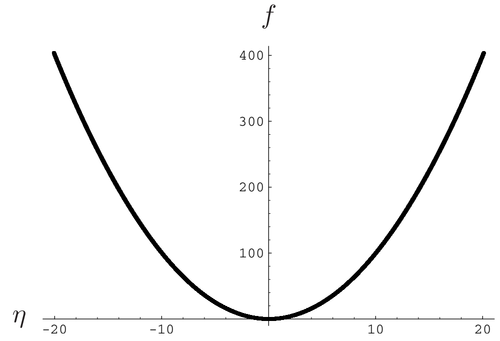


FIG. 5.3. *The solution to* (5.2) *for when the minimum is located at the origin. A step size of* $h = 0.001$ *was used and a four-term series expansion was used to find the solution through the minimum. The grains separate into two regions of equal density.*

and then the Runge–Kutta scheme is used to find the solutions from the endpoints of the series which are located at $\eta = -0.1$ and $\eta = 0.1$. In this case, the analytical solution is exactly $f = \Gamma\eta^2$ with $\Gamma = c_0^2\nu^{-2}$, which is symmetric about the origin. Therefore this case corresponds to the situation where grains separate into regions of equal density.

Another example is to find the solution when the minimum is placed at $(\eta, f) = (-2, 0)$, as shown in Figure 5.4. Again we see an asymmetric solution: to the left of the minimum f increases only to relatively small values, whereas to the right of the minimum the solution increases relatively rapidly to large values of f. We pay particular note to the way in which the curve seems to tend to a small constant for a large distance to the left of the minimum before beginning to increase.

In summary, by modifying the fundamental diagram to include a self-consistent local viscous law we indeed find that physically sensible solutions are obtained for the case of separating grains at low density, for small times and order-one viscosity, or for order-one times and small viscosity, depending on the parameters. In the solutions presented above the condition $f = 0$ (i.e., density is zero) must be satisfied at the center of the separating region. This agrees with physical intuition and also fits well with results from the idealized inviscid case. Furthermore, the solutions can be asymmetric about the origin. This corresponds to grains moving apart, possibly at different speeds, into regions of differing density. It is important that the con-

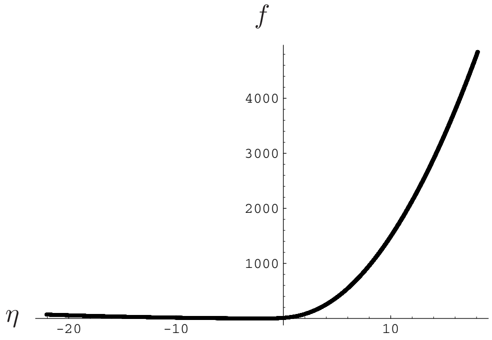


FIG. 5.4. *The solution to* (5.2) *when the minimum is located at* $(\eta, f) = (-2, 0)$*. Again we see an asymmetric solution: to the left of the minimum* f *increases only to relatively small values, whereas to the right the solution rapidly increases to large values of* f*. The solution appears to have a large region where* f *remains constant. For the above solution a step size of* $h = 0.001$ *was again used, as was a four-term series expansion through the minimum. Thus the creation of a void in a separating region seems to be described.*

tinuum model allows this kind of solution as regions of differing density are seen in direct numerical simulations of the chute flow, where clusters of different sizes develop [5]. We also found above that a large region of low density can evolve between two separating regions; see Figure 5.4. Again, one would expect this to occur for some cases of separating regions, and we also mention that numerical simulations have some large regions that are devoid of grains [5]. Finally, an origin shift of the center of the separating region is possible; i.e., the grains do not have to separate about a fixed point, which is also a physically sensible result. Consequently, it seems reasonable to conclude for cases of separating grains that the $q \sim \rho^{\frac{3}{2}}$ law from the inviscid case should be changed locally to a viscous $q \sim \rho^{\frac{1}{2}}$ law. Similar results describing clashing of grains apply at the high-density cusp $\rho \to \rho_M$ [5]. Although this fundamental diagram cannot yet be entirely supported by physical arguments, physically sensible results are nevertheless obtained. The modifications therefore appear to be admissible.

6. Remarks on steady states. We noted at the end of section 3 that general steady states do not exist for the inviscid model, except for ρ, q constant. This is not the case for the viscous model, where steady states are possible in the sense of being a large-time limit. The steady continuum equation is, from (4.1),

$$q_x = \nu \rho_{xx}. \tag{6.1}$$

Hence

$$\frac{x - c_2}{\nu} = \int_{\rho_1}^{\rho} \frac{d\rho}{q + c_1}, \tag{6.2}$$

where c_1, c_2 are constants of integration. Here for a finite integral we must choose $c_1 = 0$, and then the local behavior $q \sim \pm\rho^{\frac{1}{2}}$ from section 5 results in $\rho \sim (x - c_2)^2$ on integration of (6.2). This confirms (5.1a), which allows for a maximum or a minimum near crossover at $x = c_2$. Thus steady viscous states exist with two-way viscous flow.

The contrasting predictions from inviscid theory and viscous theory concerning steady states are due to there being a long viscous time scale in the viscous regime for small ν; evolution over times t of order $\frac{1}{\nu}$ is inferred directly from (4.1) for x

remaining of order unity. This conclusion is reinforced by the case of near-uniform density flow where $\rho = \bar{\rho} + \tilde{\rho}$, with $\bar{\rho}$ being a constant and $\tilde{\rho}$ being small. For then $c \approx \bar{c}$ is near-uniform and (4.1) reduces to

$$\tilde{\rho}_t + \bar{c}\tilde{\rho}_x = \nu\tilde{\rho}_{xx}, \tag{6.3}$$

admitting a solution of the form $\tilde{\rho} = \tilde{f}\,(x - \bar{c}t, t)$, where $\tilde{f}$ is controlled by the diffusion equation acting over the time scale of order $\frac{1}{\nu}$.

7. Further comments. If the fundamental diagram proposed in the present paper is taken to give the flux-density law for a chute flow of grains, the previous sections show that the inherent discontinuities in the model can apparently describe the formation of clusters, voids, and sudden jumps in the density during collisions and separations. Hence we can construct descriptions of significant parts of a chute flow including regions of colliding or separating grains. However, we do not claim that the fundamental diagram must describe the *entire* chute flow. The approach has also been an empirical one: we have seen that the results can describe some situations seen on chutes in reality, although the original physical arguments per se may remain open to question.

In section 3 on the inviscid model, solutions with discontinuous input were found, these being cases that yield relatively easily to analysis. Several examples of increasing complexity illustrated how shocks and fans can be used to model clusters and voids. It was noted that steady state solutions in general do not exist, except for ρ, q constant. In the future, it may be possible to include an analysis of the effects of the chute wall in an inviscid case.

In section 4, the number of solutions was extended to other cases. To obviate problems associated with the discontinuities present in the model, an artificial viscous dissipation term was added so that the governing equation is parabolic. We focused there solely on problems in which grain movement is unidirectional so that no branch-switching occurred. Hence a finite difference scheme was used to obtain numerical solutions that imitate those found in the inviscid model.

Finally, in sections 5 and 6, we showed that the work can be extended further to encompass a two-way flow in which grains can move to and fro. In order to find such solutions, the fundamental diagram has to be modified so its curvature is convex outwards at the endpoints. It was determined that such an alteration allows physically reasonable descriptions of separating and clashing grains to develop. Again, it is interesting, as section 6 pointed out, that steady state solutions do exist for the viscous model.

If the viscosity ν is small, then in general the viscous effects also are small. Exceptions are (a) shocks, where ν has a smoothing effect, as is well known [17]; (b) crossovers, which allow two-way solutions, and which are a new feature; and (c) steady states, which can exist only if $\nu \neq 0$ and appear over a long time scale, which is also a new observation.

It may be possible in future work to determine the fundamental diagram more rigorously from a direct numerical simulation. A small control volume can be considered in the computational domain and at random times a measurement of the grain density and flux within the control volume could be taken. This would be repeated for many computational runs and a distribution of points of q versus ρ produced. These points could then be collapsed onto a curve by an appropriate statistical method and a fundamental diagram would be produced. This is similar to the recent approach of Armbruster et al. [3].

We also aim to obtain two-way viscous solutions across the entire chute, involving the whole multivalued fundamental diagram. We hope to include such solutions, alongside aspects addressing other mathematical properties of the viscous system, in a future paper.

Acknowledgments. The authors acknowledge the support, financial and otherwise, of Sortex Ltd., and especially mention Dr. Gabriel Hamid and Dr. Mark Honeywood. They would also like to thank Dr. Tim Boxer of the Smith Institute for Industrial Mathematics and Systems Engineering. Useful discussions were also held with Prof. Ben Heydecker of the UCL Centre for Transport Studies.

REFERENCES

[1] I. S. Aranson and L. S. Tsimring, *Continuum description of avalanches in granular media*, Phys. Rev. E, 64 (2003), article 020301.

[2] I. S. Aranson and L. .S. Tsimring, *Continuum theory of partially fluidized granular flows*, Phys. Rev. E, 65 (2003), article 061303.

[3] D. Armbruster, D. E. Marthaler, C. Ringhofer, K. Kempf, and T.-C. Jo, *A continuum model for a re-entrant factory*, Oper. Res., 54 (2006), pp. 933–970.

[4] C. M. Dafermos, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[5] A. S. Ellis, *Modelling Chute Delivery of Grains in a Food-Sorting Process*, Ph.D thesis, University of London, 2007.

[6] E. L. Grossman, T. Zhou, and E. Ben-Naim, *Towards granular hydrodynamics in two dimensions*, Phys. Rev. E, 55 (1997), pp. 4200–4206.

[7] R. L. Hughes, *A continuum theory for the flow of pedestrians*, Transp. Res. Part B, 36 (2007), pp. 507–535.

[8] R. L. Hughes, *The flow of human crowds*, Ann. Rev. Fluid Mech., 35 (2003), pp. 169–182.

[9] J. T. Jenkins and S. B. Savage, *A theory for the rapid flow of identical, smooth, nearly elastic, spherical particles*, J. Fluid Mech., 130 (1983), pp. 187–202.

[10] J. P. Lebacque and M. M. Khoshyaran, *First order macroscopic traffic flow models: Intersection modeling, network modeling*, in Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT16), H. S. Mahmassani, ed., Hermes, 2005, pp. 365–386.

[11] M. J. Lighthill and G. B. Whitham, *On kinematic waves* II. *A theory of traffic flow on long crowded roads*, Proc. Roy. Soc. London. Ser. A, 229 (1955), pp. 317–345.

[12] J. V. Morgan, *Numerical Methods Macroscopic Traffic Models*, Ph.D. Thesis, University of Reading, 2002.

[13] K. Nagel, *Particle hopping versus fluid-dynamical models for traffic flow*, in Workshop on Traffic and Granular Flow, D. E. Wolf, M. Schreckenberg, and A. Bachem, eds., World Scientific, River Edge, NJ, 1995, pp. 41–57.

[14] J. Rajchenbach, *Granular flows*, Adv. Phys., 49 (2000), pp. 229–256.

[15] F. T. Smith and A. S. Ellis, *Air effects between interacting falling grains*, in preparation.

[16] P. Westwood, *Flow from Ejector Nozzle Arrays for Sorting Machines*, Ph.D thesis, University of London, 2005.

[17] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

[18] P. Wilson, *On the Core Flow and Turbulent Boundary Layer in a Curved Duct*, Ph.D thesis, University of London, 2003.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 330–347

OPTICAL FIBER DRAWING AND DOPANT TRANSPORT[*]

H. HUANG[†], R. M. MIURA[‡], AND J. J. WYLIE[§]

Abstract. Optical fibers are made of glass with different refractive indices in the (inner) core and the (outer) cladding regions. The difference in refractive indices arises due to a rapid transition in the concentration of a dopant across the boundary between these two regions. Fibers are normally drawn from a heated glass preform, and the different dopant concentrations in the two regions will change due to dopant diffusion and convective transport induced by the flow. In this paper, we analyze a mathematical model for the dynamics of dopant concentration changes during the fiber drawing process. Using a long-wave approximation, we show that the governing equations can be reduced to a simple diffusion equation. As a result, we are able to identify key dimensionless parameters that contribute to the diffusion process. We also derive asymptotic solutions for the temperature, cross-sectional area, and effective diffusion coefficient when there are strong temperature dependencies in the viscosity and the diffusion coefficient. Our simplified model and asymptotic solutions reduce the need for extensive numerical simulations and can be used to devise control strategies to limit excess dopant diffusion.

Key words. dopant diffusion, optical fiber drawing, incompressible viscous flow, long-wave approximation, asymptotic approximation

AMS subject classifications. 76D99, 76R50, 41A60, 76D27

DOI. 10.1137/070700176

1. Introduction. Optical fibers are drawn from a heated glass preform using mechanical pullers. The glass preform is fabricated so that there is a difference in the refractive index between the fiber core and the outer cladding region. This refractive index difference is achieved normally by adding a dopant to the inner core region [4]. Typically, dopant materials, such as oxides of germanium, phosphorus, and boron, are deposited in pure silica in the perform. However, during drawing, splicing, and fusion, the refractive index may change due to diffusion of the dopant in the glass [7, 11].

Compared to dopant concentration changes due to splicing and fusion of these fibers, dopant concentration changes during fiber drawing would appear to be more complicated since dopant diffusion depends not only on temperature, but also on a number of other factors, including the mechanics of the drawing process. In Lyytikäinen et al. [5], numerical simulations and an experimental study of specialized fibers have been carried out. For relatively low drawing speeds, it was shown that diffusion can cause a small but visible spreading of the dopant. It also was shown that larger drawing speeds and lower furnace temperatures both reduce the diffusion of dopant. Their simulations ignored advection of dopant by the flow and were based

[*] Received by the editors August 14, 2007; accepted for publication (in revised form) July 23, 2008; published electronically October 29, 2008.
http://www.siam.org/journals/siap/69-2/70017.html

[†] Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3 Canada (hhuang@yorku.ca). This author was supported by NSERC and MITACS of Canada and by the Programme of Introducing Talents of Discipline to Universities of China (18 B08018).

[‡] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (miura@njit.edu). This author was supported by the Division of Mathematical Sciences at NSF (0709092).

[§] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (mawylie@cityu.edu.hk). This author was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 103207).

on a simple diffusion equation with only a radial component. Their experiments and simulations compare favorably, which indicates that the key mechanism is captured by simple radial diffusion; however, it is not clear why this is the case and whether this assumption will hold for other parameter values. On the other hand, Yan and Pitchumani [11] carried out a full numerical simulation of the drawing process, including dopant diffusion. Although the fiber surface is a free boundary, they simplified their computations by using a prescribed surface boundary based on previous numerical studies of the drawing process by Lee and Jaluria [6]. Contrary to the conclusion in Lyytikäinen et al. [5], the numerical simulations by Yan and Pitchumani [11] show that a significant amount of diffusion occurs during the drawing process, in spite of a similar drawing environment and higher drawing speeds, which should reduce the amount of diffusion.

In this paper, we analyze a mathematical model for dopant concentration changes during optical fiber drawing. The main objective of the paper is to understand the mechanism of dopant transport during drawing. We also are interested in exploring different ways to control dopant diffusion since from a practical point of view it is desirable to minimize its effect. Based on an asymptotic analysis of this model, we are able to show that the diffusion of dopant is governed by a simple diffusion equation with only a radial component, as used by Lyytikäinen et al. [5]. However, the molecular diffusion coefficient used in [5] must be replaced by an effective diffusion coefficient, which includes a “history” factor. For typical parameter values, we show that the effective diffusion coefficient is determined mainly by two dimensionless parameters, namely, the Péclet number based on the diffusion coefficient for dopant, and a parameter that quantifies the heating strength. For large changes in viscosity, we derive analytical expressions that are uniformly valid asymptotic expansions for velocity, radius, and temperature. This allows us to find simple expressions for the effective diffusion coefficient that clearly show the way in which all of the parameters affect the diffusion process and hence reduces the need for extensive numerical simulations.

The paper is organized as follows. In section 2, the mathematical model for glass optical fiber drawing is given and subsequently simplified. We derive explicit approximations for the temperature and cross-sectional area in section 3. For these approximations, we considered two cases, one with cooling and one without cooling. In section 4, we derive asymptotic approximations for the effective diffusion in the case of no cooling.

2. Problem description. In a typical setup for glass optical fiber drawing, a cylindrical preform with radius R_0 and temperature T_0 is extruded from an input nozzle into a heating and cooling device with speed u_0; see Figure 1. At a distance L from the input nozzle, the fiber is pulled out of the device by a roller. Between the input nozzle and a distance $L_f < L$, the fiber is inside a furnace and is subjected to heating. This heating causes the viscosity of the glass to dramatically decrease, and thus facilitates rapid stretching of the fiber with moderate forces. Between the end of the furnace and the roller, the fiber is cooled by natural and forced cooling. At the nozzle input, the dopant concentration $c = c_0(r)$ will be assumed to be a given function of the radial distance from the fiber axis, r. The aim of this study is to understand how the heating, cooling, and the stretching process, as well as the diffusion and advection of the dopant, affect the dopant concentration profile when the fiber exits the device. We note that throughout this paper, subscripts 0, f, and c refer to quantities associated with the input nozzle, furnace heating, and cooling, respectively.

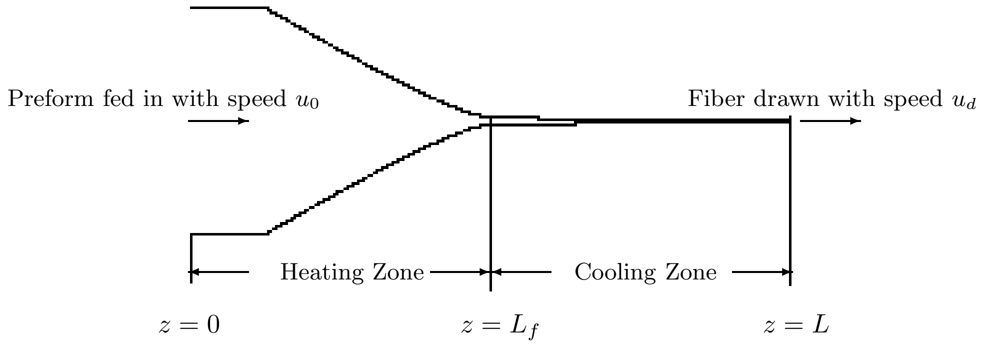


FIG. 1. *Schematic of heating and cooling zones.*

2.1. Mathematical model. We assume that the glass fiber is an incompressible fluid with temperature-dependent viscosity. Also, we assume that the dopant concentration has a negligible effect on the density, viscosity, and conductivity of the glass, the fiber remains axisymmetric, and the drawing conditions are in a steady state. Under these assumptions, the governing equations for mass, momentum, energy, and dopant concentration are given by [11]

$$\frac{\partial(\rho u)}{\partial z} + \frac{1}{r}\frac{\partial(r\rho v)}{\partial r} = 0, \tag{2.1}$$

$$\frac{\partial(\rho u^2)}{\partial z} + \frac{1}{r}\frac{\partial(r\rho uv)}{\partial r} = -\frac{\partial p}{\partial z} + 2\frac{\partial}{\partial z}\left(\mu\frac{\partial u}{\partial z}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left[r\mu\left(\frac{\partial u}{\partial r} + \frac{\partial v}{\partial z}\right)\right], \tag{2.2}$$

$$\frac{\partial(\rho uv)}{\partial z} + \frac{1}{r}\frac{\partial(r\rho v^2)}{\partial r} = -\frac{\partial p}{\partial r} + \frac{\partial}{\partial z}\left[\mu\left(\frac{\partial u}{\partial r} + \frac{\partial v}{\partial z}\right)\right] + \frac{2}{r}\frac{\partial}{\partial r}\left(r\mu\frac{\partial v}{\partial r}\right) - \frac{2\mu v}{r^2}, \tag{2.3}$$

$$\frac{\partial(\rho c_p uT)}{\partial z} + \frac{1}{r}\frac{\partial(r\rho c_p vT)}{\partial r} = \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(rk\frac{\partial T}{\partial r}\right), \tag{2.4}$$

$$\frac{\partial(uc)}{\partial z} + \frac{1}{r}\frac{\partial(rvc)}{\partial r} = \frac{\partial}{\partial z}\left(D\frac{\partial c}{\partial z}\right) + \frac{1}{r}\frac{\partial}{\partial r}\left(rD\frac{\partial c}{\partial r}\right), \tag{2.5}$$

where z is the distance from the input nozzle measured along the axis of the fiber, u and v are the velocity components in the axial and radial directions, respectively, p is the pressure, T is the temperature, c is the dopant concentration, ρ is the density, μ is the viscosity, c_p is the specific heat, $k = k_T + k_R$ is the effective conductivity [11] where k_T is the (molecular) thermal conductivity and k_R is the radiative conductivity, and D is the molecular diffusivity of the dopant. We observe that the mass, momentum, and energy equations decouple from the dopant equation.

The boundary conditions at the inlet of the furnace are

$$u = u_0,\ v = 0,\ T = T_0,\ c = c_0(r),\ 0 \le r \le R_0, \quad \text{at} \quad z = 0. \tag{2.6}$$

At a fixed downstream location, we assume the velocity is known:

$$u = u_d \quad \text{at} \quad z = L. \tag{2.7}$$

At this downstream location, boundary conditions also are needed for the radial component of the velocity v, the temperature T, and the dopant concentration c. However, we will show that in the asymptotic limit of the long-wave approximation, such boundary conditions do not play a significant role outside of a thin region near this boundary. The lateral fiber surface $r = R(z)$ is a free boundary at which the following dynamic and kinematic conditions must be applied:

$$\mathbf{n}^T \cdot \sigma \cdot \mathbf{n} = \Gamma\kappa, \; \mathbf{t}^T \cdot \sigma \cdot \mathbf{t} = 0, \; v = R'u \tag{2.8}$$

where the prime denotes differentiation with respect to z, σ is the stress tensor, $\mathbf{n} = [(1 + R'^2)^{-1/2}, \; R'(1 + R'^2)^{-1/2}]^T$ is the outward normal vector to the glass surface, $\mathbf{t} = [-R'(1 + R'^2)^{-1/2}, \; (1 + R'^2)^{-1/2}]^T$ is the corresponding vector in the tangential direction, κ is the mean curvature, and Γ is the surface tension coefficient.

The boundary condition for temperature at the fiber surface depends on whether the fiber is inside or outside of the furnace. In Lee and Jaluria [6], the heat flux q is specified when the fiber is inside the furnace. In general, q depends on many factors, such as the furnace wall temperature profile, inert gas flow, and the dimensions of the furnace, as well as the fiber temperature. Here, as in previous studies [6], we have adopted the standard Newton's cooling law[1]

$$-k\frac{\partial T}{\partial n} = q := \begin{cases} h_f(T_f - T), & 0 \le z < L_f, \\ -h_c(T - T_c), & L_f \le z \le L, \end{cases} \tag{2.9}$$

where T_f and T_c are the furnace and background temperatures, respectively, and h_f and h_c are the heat transfer coefficients for the heating from the furnace and cooling to the background, respectively. For simplicity, we will assume that the background temperature is the same as the temperature at the nozzle input, that is, $T_c = T_0$, although generalization is straightforward.

Finally, the dopant concentration satisfies the no-flux boundary condition

$$D\frac{\partial c}{\partial n} = 0 \tag{2.10}$$

at the fiber surface $r = R(z)$ and the regularity condition

$$\frac{\partial c}{\partial r} = 0 \tag{2.11}$$

at the axis of the fiber $r = 0$.

For the glass fibers used in typical optical fiber fabrication, the viscosity of the fiber varies rapidly with temperature. This rapid variation plays a fundamental role in controlling the dynamics. Empirical data for glass (see [8, 3] and the references therein) show that the viscosity can be well approximated by an Arrhenius formula or an exponential law. In this paper, we will use the exponential law in the form

$$\mu(T) = \mu_0 \exp\left(-G_\mu(T - T_0)\right) \tag{2.12}$$

where μ_0 is the viscosity of the fiber at T_0 and G_μ is a constant. The diffusion coefficient for the dopant also is normally assumed to follow the Arrhenius formula

$$D(T) = D_\infty \exp\left(-\frac{G_D}{T}\right) \tag{2.13}$$

where G_D is the activation energy divided by the universal gas constant and D_∞ is the diffusion coefficient at high temperatures.

[1] Within the furnace, there is a complicated balance between radiative and convective heat processes, but it is straightforward to generalize our approach, as outlined in this paper, to heating laws for these processes, e.g., for radiative heat transfer; see [3].

TABLE 1
List of the physical parameter values.

ρ	c_p	k_T	k_R	Γ	h_f	h_c
kg/m^3	J/(K kg)	W/(m K)	W/(m K)	kg/s	W/(m^2 K)	W/(m^2 K)
2.23×10^3	7.538×10^2	1.130	1.2×10^1	3×10^{-1}	200	20

u_0	u_d	L	L_f	R_0	T_0	T_f	μ_0	G_μ	D_∞	G_D
m/s	m/s	m	m	m	K	K	kg/(m s)	K^{-1}	m$^2/s$	K
10^{-4}	1	0.5	0.1	6×10^{-3}	300	2300	10^8	2×10^{-2}	2.4×10^{-6}	3.73×10^4

2.2. Dimensional analysis. We nondimensionalize the governing equations using the following scalings:

$$\hat{z} = \frac{z}{L},\ \hat{r} = \frac{r}{R_0},\ \hat{R} = \frac{R}{R_0},\ \hat{u} = \frac{u}{u_0},\ \hat{v} = \frac{Lv}{R_0 u_0},\ \hat{p} = \frac{R_0^2 p}{\mu_0 u_0 L},$$

$$\hat{\mu} = \frac{\mu}{\mu_0},\ \theta = \frac{T - T_0}{T_f - T_0},\ \hat{c_0} = \frac{c_0}{c_0(0)},\ \hat{c} = \frac{c}{c_0(0)},\ \hat{D} = \frac{D}{D_\infty}.$$

Substitution of these scalings into (2.1)–(2.4) yields the following dimensionless parameters, which we list along with their order of magnitude estimates based on the typical parameter values [5, 6, 11] that are listed in Table 1:

$$D_r = \frac{u_d}{u_0} \approx 10^4,\ \delta = \frac{R_0}{L} \approx 10^{-2},\ \mathrm{Re} = \frac{\rho u_0 L}{3\mu_0} \approx 10^{-9},\ \lambda = \frac{\Gamma L}{3\mu_0 u_0 R_0} \approx 10^{-3},$$

$$\mathrm{Bi} = \frac{h_f R_0}{k} \approx 10^{-1},\ \alpha_\mu = G_\mu(T_f - T_0) \approx 40,\ \alpha_D = \frac{G_D}{T_f - T_0} \approx 20,$$

$$\Theta = \frac{T_0}{T_f - T_0} \approx 0.15,\ \mathcal{P} = \frac{u_0 R_0^2}{L D_\infty} \approx 3 \times 10^{-3},$$

and

$$\mathcal{H}_f = \frac{2\sqrt{\pi} h_f L}{\rho c_p u_0 R_0} \approx 350,\ \mathcal{H}_c = \frac{2\sqrt{\pi} h_c L}{\rho c_p u_0 R_0} \approx 35,\ \ell = \frac{L_f}{L} \approx 0.2.$$

Here D_r is the draw ratio, δ is the aspect ratio, Re is the Reynolds number, λ is the ratio of surface tension forces to viscous forces, and Bi is the Biot number. The parameters α_μ and α_D measure the changes in the viscosity and the diffusion coefficient as the temperature varies between its initial value and the heater temperature, respectively, Θ is the ratio of the initial temperature to the difference between the heater and initial temperatures, $\mathcal{P}$ is the Péclet number for the dopant, $\mathcal{H}_f$ and $\mathcal{H}_c$ represent the dimensionless strengths of the heating and cooling, respectively, and ℓ represents the proportion of the length of the device that is heated by the furnace. We note that the Biot number estimated here is consistent with the value cited in [6], where the heat transfer is estimated based on an estimate for the heat flux.

2.2.1. Flow and temperature equations: Simplifications. The mass, momentum, and temperature equations clearly decouple from the dopant equation, and we begin by simplifying these equations. Based on the parameter values given in Table 1, we see that δ, Re, λ, and Bi are small. Since $\delta \ll 1$, we can use the long-wave approximation. Furthermore, since $\lambda \ll 1$ and $\mathrm{Re} \ll 1$, we can ignore the inertia and surface tension terms in the momentum equations. Finally, assuming that Bi $\ll 1$, equations (2.1)–(2.4) become (dropping the hats)

$$\frac{\partial}{\partial z}(us) = 0, \tag{2.14}$$

$$\frac{1}{s}\frac{\partial}{\partial z}\left(\mu s \frac{\partial u}{\partial z}\right) = 0, \tag{2.15}$$

$$u\frac{\partial \theta}{\partial z} = \frac{\mathcal{H}_f(1-\theta)H(\ell - z) - \mathcal{H}_c \theta H(z-\ell)}{s^{1/2}}, \tag{2.16}$$

where $s = R^2,\ \mu = \exp(-\alpha_\mu \theta)$, and H is the Heaviside step function. The derivation of these equations follows the derivation of similar equations in previous work (cf. [1, 2, 9, 10]).

The boundary conditions (2.6) and (2.7) become

$$s = 1,\ u = 1,\ \theta = 0 \quad \text{at} \quad z = 0 \tag{2.17}$$

and

$$u = D_r \quad \text{at} \quad z = 1. \tag{2.18}$$

2.2.2. Dopant equation: Long-wave approximation. The long-wave approximation of the dopant concentration equation (2.5) is

$$\mathcal{P}\left(u\frac{\partial c}{\partial z} - \frac{r}{2}\frac{\partial u}{\partial z}\frac{\partial c}{\partial r}\right) = \frac{1}{r}\frac{\partial}{\partial r}\left(rD\frac{\partial c}{\partial r}\right) \tag{2.19}$$

where

$$D = \exp\left(-\frac{\alpha_D}{\theta + \Theta}\right). \tag{2.20}$$

The boundary conditions are

$$c = c_0(r),\ 0 \le r \le R_0, \quad \text{at} \quad z = 0 \tag{2.21}$$

and

$$c_r = 0 \quad \text{at} \quad r = 0 \quad \text{and} \quad r = \sqrt{s}. \tag{2.22}$$

2.2.3. Flow and temperature equations: Reduced system. From (2.14), (2.15), and the boundary conditions on u and s in (2.17) and (2.18), it is easy to verify that

$$su = 1 \quad \text{and} \quad \mu s u_z = 2F \tag{2.23}$$

where

$$F = \frac{\ln D_r}{2\int_0^1 \mu^{-1} dz}$$

is the effective pulling force that will be obtained by using the boundary condition at $z = 1$. Using the first equation in (2.23) to eliminate u in the temperature equation

(2.16) and u_z in the second equation in (2.23), we obtain a system of two coupled first-order ordinary differential equations for s and θ:

$$s_z = -\frac{2Fs}{\mu}, \tag{2.24}$$

$$\theta_z = s^{1/2}\left[\mathcal{H}_f(1-\theta)H(\ell - z) - \mathcal{H}_c\theta H(z-\ell)\right], \tag{2.25}$$

which must be solved subject to the following boundary conditions:

$$s=1,\ \theta=0 \quad \text{at} \quad z=0 \quad \text{and} \quad s=D_r^{-1} \quad \text{at} \quad z=1. \tag{2.26}$$

2.2.4. Dopant equation: Simplification. We can further simplify the equation for dopant concentration changes by defining

$$\phi(z) \equiv \int_0^z D[\theta(z')]dz' \quad \text{and} \quad \bar{\phi} \equiv \phi(1) \tag{2.27}$$

and using the coordinate transformations $\xi = r/R$ and $\tau = \phi(z)/\bar{\phi}$. The quantity

$$\mathcal{D} = \frac{\bar{\phi}}{\mathcal{P}} \tag{2.28}$$

will be called the effective diffusion coefficient, and we obtain

$$c_\tau = \frac{\mathcal{D}}{\xi}\frac{\partial}{\partial \xi}\left(\xi\frac{\partial c}{\partial \xi}\right) \tag{2.29}$$

subject to

$$c = c_0(\xi) \quad \text{at} \quad \tau = 0 \tag{2.30}$$

and

$$c_\xi = 0 \quad \text{at} \quad \xi = 0 \quad \text{and} \quad \xi = 1. \tag{2.31}$$

The exit of the entire heating and cooling device is located at $\tau = 1$.

To summarize, we have shown that dopant transport during fiber drawing is governed by a diffusion equation in the coordinates (ξ, τ), which has the form for diffusion in cylindrically symmetric heat flow. The amount of diffusion is characterized by the effective diffusion coefficient, $\mathcal{D}$, which depends on the temperature distribution inside the fiber. In principle, the fiber temperature can be obtained by solving two coupled first-order ordinary differential equations (2.24) and (2.25) numerically. The effective diffusion coefficient can then be evaluated numerically using (2.27).

Since the range of temperature variation is large, the value of the molecular diffusion coefficient varies over several orders of magnitude and the effective diffusion coefficient is mainly determined by the portion of the fiber where the temperature is high. It can be seen that if the temperature is uniformly high, i.e., $\theta =$ constant, then $\bar{\phi} = D(\theta)$ and $\mathcal{D} = D(\theta)/\mathcal{P}$. Therefore, to estimate the value of the effective diffusion coefficient, it is important to obtain an estimate of the temperature inside the fiber. In the next two sections, we show that due to special features of the setup and parameter values used in practice, approximate solutions could be obtained for the fiber temperature. In the case without cooling, solutions also can be obtained for the effective diffusion coefficient. When compared with numerical solutions, such asymptotic solutions provide valuable insights for understanding the dopant transport mechanism.

3. Asymptotic solution for θ and s. We consider two cases: first without cooling ($\ell = 1$), in which the furnace heats the entire fiber, followed by the case with cooling ($\ell < 1$), in which the furnace heats the initial portion of the fiber, whereas the remaining portion is subjected to cooling.

3.1. $\ell = 1$. Introducing the scaled effective force $\mathcal{F} = 2Fe^{\alpha_\mu}/\ln D_r$, we can rewrite the system (2.24) and (2.25) as

$$s_z = -\mathcal{F}\ln D_r s e^{-\alpha_\mu(1-\theta)}, \tag{3.1}$$

$$\theta_z = \mathcal{H}_f\sqrt{s}(1-\theta), \tag{3.2}$$

subject to the boundary conditions $s(0) = 1$ and $\theta(0) = 0$. From the above two equations, we obtain

$$\frac{ds}{d\theta} = -\frac{\mathcal{F}\ln D_r}{\mathcal{H}_f}\frac{\sqrt{s}e^{-\alpha_\mu(1-\theta)}}{1-\theta}. \tag{3.3}$$

Integrating and using the boundary conditions, we obtain

$$s = \left(1 - \frac{\mathcal{F}\ln D_r}{2\mathcal{H}_f}\left\{E_1[\alpha_\mu(1-\theta)] - E_1[\alpha_\mu]\right\}\right)^2 \tag{3.4}$$

where

$$E_1[\eta] = \int_\eta^\infty \frac{e^{-x}}{x}dx$$

is the exponential integral. Note that up to this point no approximation has been made to (3.1) and (3.2). To proceed further, we exploit the fact that viscosity varies rapidly with temperature, that is, $\alpha_\mu \gg 1$. Note that the exponential integral has the following asymptotic approximations:

$$E_1[\eta] \sim \frac{e^{-\eta}}{\eta} \text{ as } \eta \to \infty, \qquad E_1[\eta] \sim -\ln(\eta) - \gamma \text{ as } \eta \to 0$$

where $\gamma = 0.5772\ldots$ is Euler's constant.

From (3.2), we have

$$\theta_z = \mathcal{H}_f\left(1 - \frac{\mathcal{F}\ln D_r}{2\mathcal{H}_f}\left\{E_1[\alpha_\mu(1-\theta)] - E_1[\alpha_\mu]\right\}\right)(1-\theta). \tag{3.5}$$

Since $\alpha_\mu \approx 30 \gg 1$, we have that $E_1[\alpha_\mu]$ is small. Also, $E_1[\alpha_\mu(1-\theta)]$ will be small if $1-\theta \gg \alpha_\mu^{-1}$. Therefore,

$$\theta_z = \mathcal{H}_f(1-\theta), \tag{3.6}$$

which can be solved to give

$$\theta = 1 - e^{-\mathcal{H}_f z}. \tag{3.7}$$

Clearly, $1-\theta \to 0$ as z increases to 1 for $\mathcal{H}_f \gg 1$. Thus, for $z \gg \mathcal{H}_f^{-1}$, we have

$$\theta_z = \mathcal{H}_f\left(1 + \frac{\mathcal{F}\ln D_r}{2\mathcal{H}_f}\left\{\ln[\alpha_\mu(1-\theta)] + \gamma\right\}\right)(1-\theta). \tag{3.8}$$

Letting

$$\alpha_\mu(1-\theta) = e^{-\frac{2\mathcal{H}_f}{\mathcal{F}\ln D_r}}\hat{\theta},$$

we have

$$\hat{\theta}_z = -\frac{\mathcal{F}\ln D_r}{2}\left(\ln\hat{\theta}+\gamma\right)\hat{\theta}, \tag{3.9}$$

which can be integrated to obtain

$$\ln\hat{\theta} = -\gamma + C_1 e^{-\frac{\mathcal{F}\ln D_r}{2}z} \tag{3.10}$$

where C_1 is an integration constant.

Next, we match the above solution with (3.6) to obtain

$$C_1 = \gamma + \ln\alpha_\mu + \frac{2\mathcal{H}_f}{\mathcal{F}\ln D_r}. \tag{3.11}$$

Thus,

$$\hat{\theta} = \exp\left[-\gamma + \left(\gamma + \ln\alpha_\mu + \frac{2\mathcal{H}_f}{\mathcal{F}\ln D_r}\right)e^{-\frac{\mathcal{F}\ln D_r}{2}z}\right], \tag{3.12}$$

or returning to the original variable,

$$\theta = 1 - \exp\left[-\left(\gamma + \ln\alpha_\mu + \frac{2\mathcal{H}_f}{\mathcal{F}\ln D_r}\right)\left(1 - e^{-\frac{\mathcal{F}\ln D_r}{2}z}\right)\right]. \tag{3.13}$$

Given θ, we can find s using (3.4).

The scaled pulling force $\mathcal{F}$ can be obtained by substituting $z = 1$ into (3.13) and (3.4) and using the condition $s(1) = D_r^{-1}$. Using the fact that $\alpha_\mu \gg 1$ and the asymptotic properties of E_1, we obtain

$$\mathcal{F} = 1 + \frac{2}{\ln D_r}\ln\left[1 + \frac{\mathcal{F}(\ln\alpha_\mu+\gamma)\ln D_r}{2\mathcal{H}_f}\right]. \tag{3.14}$$

In general, to obtain $\mathcal{F}$, this equation must be solved numerically. But for typical parameter values, we have

$$\frac{(\ln\alpha_\mu+\gamma)\ln D_r}{2\mathcal{H}_f} \sim 10^{-1} \ll 1, \tag{3.15}$$

and so we can obtain an asymptotic estimate for $\mathcal{F}$ in closed form:

$$\mathcal{F} = 1 + \frac{\ln\alpha_\mu+\gamma}{\mathcal{H}_f}. \tag{3.16}$$

In Figure 2, we plot the asymptotic and numerical solutions, and it can be seen that the agreement between the two solutions is excellent.

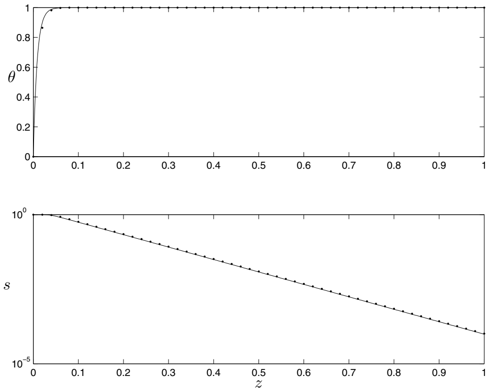


FIG. 2. *Numerical (dots) vs. asymptotic (lines) solutions. The parameter values are* $\mathcal{H}_f = 100$, $\alpha_\mu = 30$, *and* $D_r = 10^4$.

3.2. $\boldsymbol{\ell < 1}$. When there is cooling, the previous solution is valid up to the location $z = \ell$, but for $z > \ell$, the area and temperature are determined from (2.24) and (2.25):

$$s_z = -\mathcal{F} \ln(D_r) e^{\alpha_\mu(\theta-1)} s, \tag{3.17}$$

$$\theta_z = -\mathcal{H}_c \sqrt{s}\theta, \tag{3.18}$$

subject to the boundary conditions $s(\ell) = s_\ell$ and $\theta(\ell) = \theta_\ell$.

We rescale the variables using

$$\vartheta = \alpha_\mu(\theta_\ell - \theta), \quad \mathfrak{s} = \sqrt{\frac{s}{s_\ell}}, \quad y = \frac{\mathcal{F}\ln(D_r)e^{\alpha_\mu(\theta_\ell-1)}}{2}(z-\ell),$$

and (3.17) and (3.18) become

$$\mathfrak{s}_y = -\mathfrak{s}e^{-\vartheta}, \tag{3.19}$$

$$\vartheta_y = \mathcal{A}\mathfrak{s}(1-\epsilon\vartheta) \tag{3.20}$$

where

$$\mathcal{A} = \frac{2\sqrt{s_\ell}\mathcal{H}_c\alpha_\mu\theta_\ell e^{\alpha_\mu(1-\theta_\ell)}}{\mathcal{F}\ln(D_r)}, \quad \epsilon = \frac{1}{\alpha_\mu\theta_\ell} \ll 1,$$

and subject to $\vartheta(0) = 0$ and $\mathfrak{s}(0) = 1$.

From (3.19) and (3.20), we obtain

$$\frac{d\vartheta}{d\mathfrak{s}} = -\mathcal{A}(1-\epsilon\vartheta)e^{\vartheta}. \tag{3.21}$$

Using $\vartheta(0) = 0$, $\mathfrak{s}(0) = 1$, and integration by parts, we obtain

$$\mathfrak{s} = 1 - \frac{1}{\mathcal{A}}\int_0^{\vartheta}\frac{e^{-w}}{1-\epsilon w}dw = 1 + \frac{1}{\mathcal{A}}\left(\frac{e^{-\vartheta}}{1-\epsilon\vartheta} - 1\right) - \frac{\epsilon}{\mathcal{A}}\int_0^{\vartheta}\frac{e^{-w}}{(1-\epsilon w)^2}dw. \tag{3.22}$$

Given typical parameter values, one can readily see that the scaled temperature ϑ at the exit is much greater than ϵ^{-1}. Therefore, $\epsilon/(1-\epsilon\vartheta) \ll 1$, and (3.22) can be approximated as

$$\mathfrak{s} = 1 + \frac{1}{\mathcal{A}}\left(\frac{e^{-\vartheta}}{1-\epsilon\vartheta} - 1\right). \tag{3.23}$$

We now proceed with three different cases: $\mathcal{A} = 1$, $\mathcal{A} < 1$, and $\mathcal{A} > 1$.

Case 1. $\boldsymbol{\mathcal{A} = 1}$. In this case, we have

$$\mathfrak{s} = e^{-\vartheta} \tag{3.24}$$

and the equation for ϑ becomes

$$\vartheta_y = e^{-\vartheta}, \tag{3.25}$$

which can be solved as

$$\vartheta = \ln(y+1) \tag{3.26}$$

using the boundary condition $\vartheta(0) = 0$. Thus, the leading-order solution for $\mathfrak{s}$ is

$$\mathfrak{s} = \frac{1}{(y+1)}. \tag{3.27}$$

Case 2. $\boldsymbol{\mathcal{A} < 1}$. Note that $\mathcal{A}$ is the scaled cooling strength, and when $\mathcal{A} < 1$, we expect that the temperature is bounded below by $\mathcal{A} = 1$. Using the solution in Case 1, we note that ϑ is a monotonically increasing function with a maximum value

$$\vartheta_{max} = \ln\left(\frac{\mathcal{F}\ln(D_r)e^{\alpha_\mu(\theta_\ell-1)}}{2}(1-\ell) + 1\right). \tag{3.28}$$

The quantity ϑ_{max} is an order one quantity as long as $\ln(\ln(D_r)) \ll \epsilon^{-1}$, so we conclude that ϑ also will be an order one quantity.

Having established that $\vartheta = O(1)$, we can approximate (3.23) by

$$\mathfrak{s} = 1 + \frac{1}{\mathcal{A}}\left(e^{-\vartheta} - 1\right). \tag{3.29}$$

From (3.20), we obtain

$$\vartheta_y = \mathcal{A}\mathfrak{s} = e^{-\vartheta} + \mathcal{A} - 1, \tag{3.30}$$

which can be integrated to give

$$\vartheta = \ln\frac{\mathcal{A}e^{(\mathcal{A}-1)y} - 1}{\mathcal{A} - 1} \tag{3.31}$$

after applying the condition $\vartheta(0) = 0$. Note that this solution also applies to the case of $\mathcal{A} = 1$ by taking the limit $\mathcal{A} \to 1$, which yields (3.26). The solution for $\mathfrak{s}$ can be obtained as

$$\mathfrak{s} = \frac{\mathcal{A} - 1}{A - e^{-(\mathcal{A}-1)y}}. \tag{3.32}$$

Case 3. $\mathcal{A} > 1$. In this case, we need to consider two scenarios: when $\vartheta = O(1)$ and when ϑ is large.

(i) $\vartheta = O(1)$. When $\vartheta = O(1)$, the analysis is identical to the case for $A < 1$, yielding the same formulas (3.31) and (3.32).

(ii) *Large* ϑ. When $1 \ll \vartheta \ll \epsilon^{-1}$ and $\mathcal{A} > 1$, the approximation of (3.23), neglecting exponentially small terms, is given by

$$\mathfrak{s} = 1 - \frac{1}{\mathcal{A}}, \tag{3.33}$$

which can be combined with (3.20) to yield

$$\vartheta_y = (\mathcal{A} - 1)(1 - \epsilon\vartheta). \tag{3.34}$$

This equation can be solved as

$$\vartheta = \frac{1 - C_2 e^{-\epsilon(\mathcal{A}-1)y}}{\epsilon} \tag{3.35}$$

where C_2 is an integration constant. Since we obtained the solution based on the assumption that ϑ is not small, we cannot use the boundary condition $\vartheta(0) = 0$. To determine C_2, we need to match the small ϑ solution given by (3.31) with that for large ϑ given by (3.35). Taking the limit of $\vartheta \to 0$ from (3.35), we obtain

$$\vartheta \sim \frac{1 - C_2(1 - \epsilon(\mathcal{A} - 1)y)}{\epsilon}.$$

For large ϑ, i.e., $y \to 0$, (3.31) gives

$$\vartheta \sim (\mathcal{A} - 1)y + \ln \frac{\mathcal{A}}{\mathcal{A} - 1}.$$

Comparing the two expressions, the only choice we have is

$$C_2 = 1 - \epsilon \ln \frac{\mathcal{A}}{\mathcal{A} - 1},$$

and the solution becomes

$$\vartheta = \frac{1 - e^{-\epsilon(\mathcal{A}-1)y}}{\epsilon} + \left(\ln \frac{\mathcal{A}}{\mathcal{A} - 1}\right) e^{-\epsilon(\mathcal{A}-1)y}. \tag{3.36}$$

Note that when $\mathcal{A} \gg 1$, $C_2 \sim 1$ and ϑ can be approximated by

$$\vartheta = \frac{1 - e^{-\epsilon(\mathcal{A}-1)y}}{\epsilon}. \tag{3.37}$$

(iii) *Uniformly valid solution.* The uniformly valid solution for ϑ obtained by combining the solutions for small and large values of ϑ is given by

$$\begin{aligned}\vartheta = \ln \frac{\mathcal{A}e^{(\mathcal{A}-1)y} - 1}{\mathcal{A} - 1} + \frac{1 - e^{-\epsilon(\mathcal{A}-1)y}}{\epsilon} + \ln \frac{\mathcal{A}}{\mathcal{A} - 1} e^{-\epsilon(\mathcal{A}-1)y} \\ - (\mathcal{A} - 1)y - \ln \frac{\mathcal{A}}{\mathcal{A} - 1}.\end{aligned} \tag{3.38}$$

The solution for $\mathfrak{s}$ can be obtained using (3.23).

Solutions in original variables. Using the original variables and $\theta = \theta_\ell - \alpha_\mu^{-1}\vartheta$, the temperature solution is given by

$$\theta = \theta_\ell - \frac{1}{\alpha_\mu} \ln \frac{\mathcal{A} \exp\left[(\mathcal{A}-1)\mathcal{B}(z-\ell)\right] - 1}{\mathcal{A}-1} \tag{3.39}$$

for $\mathcal{A} \leq 1$ and

$$\begin{aligned}\theta = \theta_\ell &- \frac{1}{\alpha_\mu} \ln \frac{\mathcal{A} \exp\left[(\mathcal{A}-1)\mathcal{B}(z-\ell)\right] - 1}{\mathcal{A}-1} \\ &- \left(\theta_\ell - \frac{1}{\alpha_\mu} \ln \frac{\mathcal{A}}{\mathcal{A}-1}\right)\left(1 - \exp\left[-\frac{(\mathcal{A}-1)\mathcal{B}}{\alpha_\mu \theta_\ell}(z-\ell)\right]\right) + \frac{1}{\alpha_\mu}(\mathcal{A}-1)\mathcal{B}(z-\ell)\end{aligned} \tag{3.40}$$

for $\mathcal{A} > 1$. Here $\mathcal{B} = \mathcal{F} \ln(D_r) e^{\alpha_\mu(\theta_\ell - 1)}/2$.

In both cases, the solution for the cross-sectional area in the original variables is obtained using (3.23):

$$s = s_\ell \left(1 - \frac{1}{\mathcal{A}} + \frac{\theta_\ell \exp[\alpha_\mu(\theta - \theta_\ell)]}{\mathcal{A}\theta}\right)^2. \tag{3.41}$$

Using $s(1) = D_r^{-1}$, we can obtain $\mathcal{F}$ (and F) by solving the following equation numerically:

$$D_r^{-1} = s_\ell \left(1 - \frac{1}{\mathcal{A}} + \frac{\theta_\ell \exp[\alpha_\mu(\theta(1) - \theta_\ell)]}{\mathcal{A}\theta(1)}\right)^2 \tag{3.42}$$

where $\theta(1)$ is evaluated at $z = 1$ using (3.39) or (3.40), depending on the value of $\mathcal{A}$.

In Figures 3 and 4, we have plotted both the numerical and the asymptotic solutions for various values of the draw ratio D_r and for two different values of the cooling parameter $\mathcal{H}_c$, respectively. It can be seen that the solutions agree very well with each other. Therefore, instead of relying on numerical solutions, one can use (3.39) and (3.40) to evaluate the effective diffusion coefficient for the dopant. This will be shown in the next section.

4. Dopant diffusion.

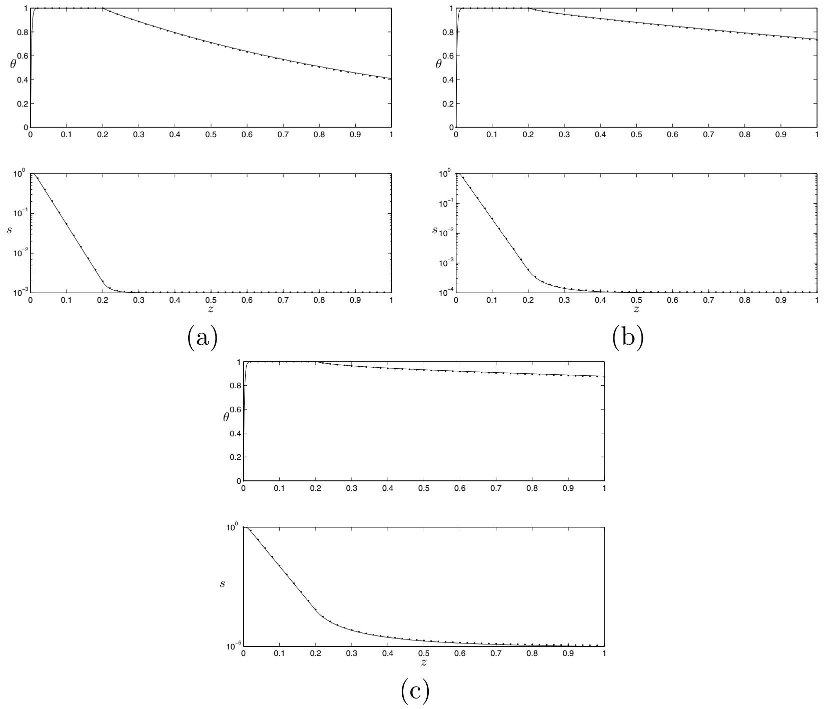
Since dopant transport follows a standard diffusion equation, the solution is completely determined by the effective diffusion coefficient. When the diffusion coefficient is small (as in the case of dopant diffusion), the effect of the boundary on the dopant distribution is also small. Therefore, we can ignore the boundary and solve (2.29) on an infinite domain as an approximation.[2] In this case, the solution for the dopant concentration can be written as

$$c(\tau, \xi) = 2\pi \int_0^1 G(\tau, \xi; \eta) c_0(\eta) \eta d\eta. \tag{4.1}$$

[2] If we consider the boundary effect, we can use the series solution given by

$$c(\tau, \xi) = 2\left[1 + \frac{r_*}{R} \sum_{m=1}^{\infty} e^{-\lambda_m^2 \mathcal{D}\tau} \frac{J_0(\lambda_m \xi) J_1(\lambda_m r_* R^{-1})}{\lambda_m J_0(\lambda_m)^2}\right]$$

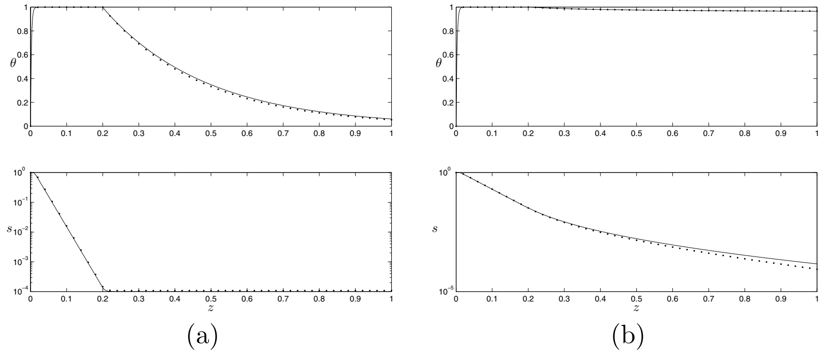
where λ_m are the zeros of $J_1(\lambda)$. As long as the diffusion is not too small, only a small number of terms is needed. For small diffusion coefficient values, convergence becomes slow, and we can switch to the solution based on the infinite domain Green's function.



(a) (b)

(c)

FIG. 3. *Numerical (dots) vs. asymptotic solutions (solid line) given by* (3.40) *and* (3.41)*:* (a) $D_r = 10^3$ *($\mathcal{A} = 3.67$);* (b) $D_r = 10^4$ *($\mathcal{A} = 1.71$); and* (c) $D_r = 10^5$ *($\mathcal{A} = 1.20$). Values of the other parameters are $\mathcal{H} = 350$, $\mathcal{H}_c = 35$, $\alpha_\mu = 40$, $\ell = 0.2$.*



(a) (b)

FIG. 4. *Numerical (dots) vs. asymptotic solutions (solid line) given by* (3.39), (3.40)*, and* (3.41)*:* (a) $\mathcal{H}_c = 350$ *($\mathcal{A} = 6.96$); and* (b) $\mathcal{H}_c = 1$ *($\mathcal{A} = 0.78$). Values of the other parameters are $\mathcal{H} = 350$, $D_r = 10^4$, $\alpha_\mu = 40$, $\ell = 0.2$.*

If we know the value of $\mathcal{D}$, then the Green's function is given by

$$G(\tau, \xi; \eta) = \frac{1}{4\pi \mathcal{D}\tau} \exp\left(-\frac{\xi^2 + \eta^2}{4\mathcal{D}\tau}\right) I_0\left(\frac{\xi\eta}{2\mathcal{D}\tau}\right) \tag{4.2}$$

where I_0 is the modified Bessel function of the first kind.

In order to illustrate how the effective diffusion coefficient $\mathcal{D}$ is affected by various parameter values, we obtain an asymptotic approximation for the case with no cooling

($\ell = 1$). Recall that in this case, the temperature is given by (3.13):

$$\theta = 1 - \exp\left[-C_1\left(1 - e^{-\frac{\mathcal{F}\ln D_r}{2}z}\right)\right],$$

with C_1 given by (3.11) and the effective diffusion coefficient is given by $\mathcal{D} = \bar{\phi}/\mathcal{P}$ where $\bar{\phi}$ is given by (2.27) with D given by (2.20).

Introducing new variables $\zeta = 1 - \exp\left(-\frac{\mathcal{F}\ln D_r}{2}z\right)$ and $\varphi(\zeta) \equiv \phi(z)$, we have

$$d\zeta = \frac{\mathcal{F}\ln D_r}{2}(1-\zeta)dz$$

and

$$\varphi_\zeta = \frac{2}{\mathcal{F}\ln D_r}\frac{\exp\left(-\frac{\alpha_D}{\Theta+1-\exp(-C_1\zeta)}\right)}{1-\zeta}, \quad \varphi(0) = 0. \tag{4.3}$$

Since $C_1 \gg 1$, the solution can be found for two cases: $\zeta \sim C_1^{-1}$ and $\zeta \sim 1$.

Case I. $\boldsymbol{\zeta \sim C_1^{-1} \ll 1}$. In this case, we use a new variable $\hat{\zeta} = C_1\zeta$ and denote the solution by $\varphi^{(i)}$ (inner solution) which satisfies

$$\varphi^{(i)}_{\hat{\zeta}} = \frac{2\exp\left(-\frac{\alpha_D}{\Theta+1-\exp(-\hat{\zeta})}\right)}{C_1\mathcal{F}\ln D_r}, \quad \varphi^{(i)}(0) = 0. \tag{4.4}$$

The inner solution can be obtained as

$$\varphi^{(i)}(\hat{\zeta}) = -\frac{2}{C_1\mathcal{F}\ln D_r}\left[E_1(x) - \exp\left(-\frac{\alpha_D}{\Theta+1}\right)E_1\left(x - \frac{\alpha_D}{\Theta+1}\right)\right]_{x=\alpha_D/\Theta}^{x=\alpha_D/(\Theta+1-\exp(-\hat{\zeta}))}. \tag{4.5}$$

Case II. $\boldsymbol{\zeta \sim 1}$. In this case, we denote the solution by $\varphi^{(o)}$ (outer solution) which satisfies

$$\varphi^{(o)}_\zeta = \frac{2}{\mathcal{F}\ln D_r}\frac{\exp(-\frac{\alpha_D}{\Theta+1})}{1-\zeta}. \tag{4.6}$$

The solution can be obtained as

$$\varphi^{(o)}(\zeta) = -\frac{2\exp(-\frac{\alpha_D}{\Theta+1})}{\mathcal{F}\ln D_r}\ln(1-\zeta) + C_3 \tag{4.7}$$

where C_3 is an integration constant.

Matching. To determine the integration constant C_3 and the outer solution, $\varphi^{(o)}$, we need to match the two solutions as follows:

$$\lim_{\hat{\zeta}\to\infty}\varphi^{(i)}(\hat{\zeta}) = \lim_{\zeta\to 0}\varphi^{(o)}(\zeta).$$

Since

$$\varphi^{(o)}(\zeta) \sim C_3 + \frac{2\exp(-\frac{\alpha_D}{\Theta+1})}{\mathcal{F}\ln D_r}\zeta \quad \text{as} \quad \zeta \to 0 \tag{4.8}$$

and $E_1[z] \sim -\gamma - \ln(z)$ for small z, we have

$$\begin{aligned}
\varphi^{(i)}(\hat{\zeta}) \sim & -\frac{2}{C_1 \mathcal{F} \ln D_r} \Bigg\{ E_1\left(\frac{\alpha_D}{\Theta+1}\right) - E_1\left(\frac{\alpha_D}{\Theta}\right) - \exp\left(-\frac{\alpha_D}{\Theta+1}\right) \\
& \times \left[-\gamma - \ln\left(\frac{\alpha_D}{\Theta+1-\exp\left(-\hat{\zeta}\right)} - \frac{\alpha_D}{\Theta+1}\right)\right] \\
& + \exp\left(-\frac{\alpha_D}{\Theta+1}\right) E_1\left(\frac{\alpha_D}{\Theta} - \frac{\alpha_D}{\Theta+1}\right) \Bigg\} \\
\approx & -\frac{2}{C_1 \mathcal{F} \ln D_r} \left\{ E_1\left(\frac{\alpha_D}{\Theta+1}\right) - E_1\left(\frac{\alpha_D}{\Theta}\right) - \exp\left(-\frac{\alpha_D}{\Theta+1}\right) \right. \\
& \left. \times \left[-\gamma - \ln\alpha_D + \hat{\zeta} + 2\ln(\Theta+1)\right] + \exp\left(-\frac{\alpha_D}{\Theta+1}\right) E_1\left(\frac{\alpha_D}{\Theta(\Theta+1)}\right) \right\} \\
\approx & -\frac{2}{C_1 \mathcal{F} \ln D_r} \left\{ E_1\left(\frac{\alpha_D}{\Theta+1}\right) - E_1\left(\frac{\alpha_D}{\Theta}\right) - \exp\left(-\frac{\alpha_D}{\Theta+1}\right) \right. \\
& \left. \times [-\gamma - \ln\alpha_D + 2\ln(\Theta+1)] + \exp\left(-\frac{\alpha_D}{\Theta+1}\right) E_1\left(\frac{\alpha_D}{\Theta(\Theta+1)}\right) \right\} \\
& + \frac{2\exp\left(-\frac{\alpha_D}{\Theta+1}\right)}{\mathcal{F}\ln D_r} \frac{\hat{\zeta}}{C_1} \quad \text{as} \quad \hat{\zeta} \to \infty.
\end{aligned} \tag{4.9}$$

We obtain

$$\begin{aligned}
C_3 = & -\frac{2}{C_1 \mathcal{F} \ln D_r} \left\{ E_1\left(\frac{\alpha_D}{\Theta+1}\right) - E_1\left(\frac{\alpha_D}{\Theta}\right) - \exp\left(-\frac{\alpha_D}{\Theta+1}\right) \right. \\
& \left. \times [-\gamma - \ln\alpha_D + 2\ln(\Theta+1)] + \exp\left(-\frac{\alpha_D}{\Theta+1}\right) E_1\left(\frac{\alpha_D}{\Theta(\Theta+1)}\right) \right\}.
\end{aligned} \tag{4.10}$$

Using the original variable, we have

$$\phi^{(o)}(z) = \exp\left(-\frac{\alpha_D}{\Theta+1}\right) z + C_3, \tag{4.11}$$

from which we obtain the effective diffusion coefficient as

$$\mathcal{D} = \frac{\exp\left(-\frac{\alpha_D}{\Theta+1}\right) + C_3}{\mathcal{P}}. \tag{4.12}$$

First, we observe that the first term on the right-hand side of $\mathcal{D}$ is given by $\mathcal{P}^{-1}\exp\left(-\frac{\alpha_D}{\Theta+1}\right)$, which is the value of the effective diffusion coefficient when $\theta = \Theta$ for all z, i.e., uniform temperature. The correction term C_3 is inversely proportional to

$$2\mathcal{H}_f + (\ln\alpha_\mu + \gamma)\left(1 + \frac{\ln\alpha_\mu + \gamma}{\mathcal{H}_f}\right)\ln D_r.$$

The parameter $\mathcal{H}_f$ is typically much larger than the other parameters. Therefore, the above asymptotic analysis clearly shows that the effective diffusion is extremely weakly dependent on the draw ratio D_r.

On the other hand, $\mathcal{D}$ is strongly affected by the Péclet number $\mathcal{P}$. Therefore, the most effective way to control excessive diffusion is to increase $\mathcal{P}$, which can be achieved by using a relatively large feeding speed. We note that the diffusion coefficient given here is for the scaled fiber radius. Thus, the absolute length of diffusion is proportional to the initial fiber radius R_0. For illustrative purposes, we have evaluated the effective diffusion coefficient given in (4.12) using the parameter values listed in Table 1 and keeping the feeding speed as a free parameter, yielding

$$\mathcal{D} \approx \frac{3.02 \times 10^{-9}}{u_0} - \frac{1.44 \times 10^{-11}}{4.27u_0^2 + 5.10 \times 10^{-5}} \text{ m}^2/\text{s}.$$

Here the second term comes from the correction term $C_3/\mathcal{P}$. Taking the value of $u_0 = 10^{-4}$ m/s (cf. [6]), we obtain $\mathcal{D} \approx 2.99 \times 10^{-5}$ m^2/s with the correction term and 3.02×10^{-5} m^2/s without the correction term. Next we compute the effective diffusion coefficient using numerical quadrature, based on (2.27), (2.28), and (3.13), which yields $\mathcal{D} = 2.99 \times 10^{-5}$ m^2/s. For a higher feeding speed, e.g., $u_0 = 3 \times 10^{-3}$ m/s, the asymptotic solution is $\mathcal{D} \approx 8.45 \times 10^{-7}$ m^2/s with the correction term and 1.01×10^{-6} m^2/s without the correction term, while the numerical solution is $\mathcal{D} = 8.5 \times 10^{-7}$ m^2/s. In both cases the asymptotic solution is a very good approximation when it is compared with the numerical solution, especially when the correction term is included.

Finally, to find $\mathcal{D}$ for the case with cooling, we could solve for the temperature from (2.24)–(2.25) with boundary condition (2.26). The effective diffusion coefficient could be computed using the integral $\int_0^1 D(\theta)dz$. In general, we can apply numerical methods, e.g., finite difference to solve the flow and temperature equations (2.24) and (2.25) and numerical quadrature to the integral to find an approximation of $\mathcal{D}$. However, since we already obtained an asymptotic solution for the temperature, $\mathcal{D}$ can be computed easily by evaluating the integral using numerical quadrature.

5. Conclusion. In this paper, we have shown that the long-wave approximation can be used to dramatically simplify the governing equations for dopant transport in optical fiber drawing. The viscosity and diffusion coefficient vary rapidly with temperature, which makes direct numerical simulations difficult. However, we take advantage of these rapid changes to derive asymptotic approximations of the solution. We show that the transport of dopant satisfies a simple diffusion equation with an effective diffusion coefficient that can be computed easily using our asymptotic solution. Our solution shows that the feeding speed is the most effective way to control dopant diffusion from the core into the cladding region. Using our asymptotic solution, other control strategies can also be developed. For example, one can use an optimal control framework based on a cost function that maximizes fiber production and minimizes dopant diffusion and which uses the feeding and drawing speeds or the heating and cooling rates as control variables. However, such a study is beyond the scope of this paper and will be pursued in future work.

REFERENCES

[1] A. D. Fitt, K. Furusawa, T. M. Monro, and C. P. Please, *Modeling the fabrication of hollow fibers: Capillary drawing*, J. Lightwave Technol., 19 (2001), pp. 1924–1931.

[2] G. Gupta and W. W. Schultz, *Non-isothermal flows of Newtonian slender glass fibers*, Internat. J. Non-Linear Mech., 33 (1998), pp. 151–163.

[3] H. Huang, R. M. Miura, W. P. Ireland, and E. Puil, *Heat-induced stretching of a glass tube under tension: Application to glass microelectrodes*, SIAM J. Appl. Math., 63 (2003), pp. 1499–1519.

[4] T. Izawa, *Early days of VAD process*, IEEE J. Selected Topics Quantum Electronics, 6 (2000), pp. 1220–1227.

[5] K. Lyytikäinen, S. T. Huntington, A. L. G. Carter, P. McNamara, S. Fleming, J. Abramczyk, I. Kaplin, and G. Schötz, *Dopant diffusion during optical fibre drawing*, Optical Express, 12 (2004), pp. 972–977.

[6] S. H.-K. Lee and Y. Jaluria, *Effects of variable properties and viscosity dissipation during optical fiber drawing*, Trans. ASME, 118 (1996), pp. 350–358.

[7] E. Pone, X. Daxhelet, and S. Lacroix, *Refractive index profile of fused-fiber couplers cross-section*, Optical Express, 12 (2004), pp. 1036–1044.

[8] H. Scholze, *Glass, Nature, Structure, and Properties*, M. J. Lakin, trans., Springer-Verlag, New York, 1990, pp. 255–272.

[9] J. Wylie and H. Huang, *Extensional flows with viscous heating*, J. Fluid Mech., 570 (2007), pp. 359–370.

[10] J. Wylie, H. Huang, and R. M. Miura, *Thermal instability in drawing viscous threads*, J. Fluid Mech., 570 (2007), pp. 1–13.

[11] Y. Yan and R. Pitchumani, *Numerical study on the dopant concentration and refractive index profile evolution in an optical fiber manufacturing process*, Int. J. Heat Mass Transfer, 49 (2006), pp. 2097–2112.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 348–374

STATIONARY SOLUTIONS OF DRIVEN FOURTH- AND SIXTH-ORDER CAHN–HILLIARD-TYPE EQUATIONS[*]

M. D. KORZEC[†], P. L. EVANS[‡], A. MÜNCH[§], AND B. WAGNER[¶]

Abstract. New types of stationary solutions of a one-dimensional driven sixth-order Cahn–Hilliard-type equation that arises as a model for epitaxially growing nanostructures, such as quantum dots, are derived by an extension of the method of matched asymptotic expansions that retains exponentially small terms. This method yields analytical expressions for far-field behavior as well as the widths of the humps of these spatially nonmonotone solutions in the limit of small driving force strength, which is the deposition rate in case of epitaxial growth. These solutions extend the family of the monotone kink and antikink solutions. The hump spacing is related to solutions of the Lambert W function. Using phase-space analysis for the corresponding fifth-order dynamical system, we use a numerical technique that enables the efficient and accurate tracking of the solution branches, where the asymptotic solutions are used as initial input. Additionally, our approach is first demonstrated for the related but simpler driven fourth-order Cahn–Hilliard equation, also known as the convective Cahn–Hilliard equation.

Key words. convective Cahn–Hilliard, quantum dots, exponential asymptotics, matching, dynamical systems

AMS subject classifications. 34E05, 74K35, 65P99

DOI. 10.1137/070710949

1. Introduction.

A paradigm for phase separating systems such as binary alloys is the Cahn–Hilliard equation for the phase fraction u,

$$u_t + \left(Q(u) + \varepsilon^2 u_{xx}\right)_{xx} = 0, \tag{1.1}$$

where $Q(u)$ is the negative derivative of the double-well potential $-\mathcal{F}$, typically

$$Q(u) = \mathcal{F}'(u) = u - u^3. \tag{1.2}$$

The long-time dynamics are characterized by the logarithmically slow coarsening process of phases, corresponding to local minima of the potential, separated by interfaces of width ε. This process is described well by the motion of equidistantly spaced smoothed shock solutions or *kinks* ("positive kinks") and *antikinks* ("negative kinks") which connect the local minimum of $\mathcal{F}(u)$ at $u = -1$ to that at $u = 1$ and vice versa.

In recent years, an extension of this model has been studied for the case when the phase separating system is driven by an external field [16, 27]. In one space dimension it can be written as

$$u_t - \nu u u_x + \left(Q(u) + \varepsilon^2 u_{xx}\right)_{xx} = 0, \tag{1.3}$$

[*] Received by the editors December 14, 2007; accepted for publication (in revised form) June 10, 2008; published electronically November 5, 2008. This work was performed as part of Project C-10 of the DFG research center Matheon.
http://www.siam.org/journals/siap/69-2/71094.html

[†] Corresponding author. Weierstrass Institute for Applied Analysis and Stochastics (WIAS), D-10117 Berlin, Germany (korzec@wias-berlin.de).

[‡] Institute for Mathematics, Humboldt University of Berlin, D-10099 Berlin, Germany (pevans@mathematik.hu-berlin.de).

[§] School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK (andreas.muench@nottingham.ac.uk). The work of this author was partially supported by the Heisenberg Fellowship of the DFG (grant MU 1626/3).

[¶] Weierstrass Institute for Applied Analysis and Stochastics (WIAS), D-10117 Berlin, Germany (wagnerb@wias-berlin.de).

where ν denotes the strength of the external field. This equation, the convective Cahn–Hilliard (CCH) equation, also arises as a model for the evolution of the morphology of steps on crystal surfaces [21], and the growth of thermodynamically unstable crystal surfaces into a melt with kinetic undercooling and strongly anisotropic surface tension [17, 11, 9].

The dynamics of this model as $\nu \to 0$ are characterized by coarsening, as is typical for the Cahn–Hilliard equation ($\nu = 0$) [7, 26]. If $\nu \to \infty$, using the transformation $u \mapsto u/\nu$ in (1.3) one obtains the Kuramoto–Sivashinski equation, which is a well-known model for spatiotemporal chaotic dynamics (see, e.g., [10] and references therein). Recently, Eden and Kalantarov [6] also established the existence of a finite-dimensional inertial manifold for the CCH equation, viewed as an infinite-dimensional dynamical system.

A related higher order evolution equation arises in the context of epitaxially growing thin films (for a review on self-ordered nanostructures on crystal surfaces see Shchukin and Bimberg [24]). Here, the formation of *quantum dots* and their faceting has been described by the sixth-order equation

$$u_t - \nu u u_x - \big(Q(u) + \varepsilon^2 u_{xx}\big)_{xxxx} = 0, \tag{1.4}$$

where u denotes the surface slope, ν is proportional to the deposition rate [22], and $Q(u)$ is given with (1.2); it is assumed to have this form from now on throughout the paper. The high order derivatives are a result of the additional regularization energy which is required to form an edge between two plane surfaces with different orientations. This implies that the crystal surface tension also depends on curvature, which becomes very high at edges as the parameter ε goes to zero. In analogy to the Cahn–Hilliard equation, here the phases are the orientations of the facets. This *higher order convective* Cahn–Hilliard (HCCH) equation shares many properties with the CCH equation. In both cases the dynamics are described by conserved order parameters if $\nu = 0$. They also share characteristic coarsening dynamics as $\nu \to 0$ and chaotic dynamics as ν becomes large. To understand the complicated structure of the solutions it is instructive to study first the stationary solutions and their stability, as has been done for the CCH equation [28, 16]. For small ν, the stationary solutions for both equations have been characterized by the monotone *kink* and *antikink* solutions [16, 22]. Recently new spatially nonmonotone solutions were found for the lower order equation [28]. In this study we establish that the HCCH equation also possesses such nonmonotone solutions. We show this by using phase-space methods for the corresponding fifth-order boundary value problem. We use the expression "simple" or "monotone" for a solution that connects the maximal value of $u(x)$ to the minimal value without any humps on the way down, although these extrema exist and lead to nonmonotonicity even for simple (anti-)kink solutions of the HCCH equation.

Since the treatment of this high order problem is not straightforward, one part of this study is concerned with the development of an approach that accurately locates the heteroclinic connections in the five-dimensional phase space. We find that these stationary solutions develop oscillations whose width and amplitude increase as $\nu \to 0$.

In the second part of this study we derive an analytic expression for the width and amplitude within the asymptotic regime of small external field strength. For the CCH equation we find that the width has a logarithmic dependency on the strength of the external field, while for the HCCH equation our analysis yields a dependency in terms of the Lambert W function. In order to arrive at these expressions we solve the fifth-order equation by a combination of the method of matched asymptotic expansions and exponential asymptotics. We first demonstrate our approach for the third-order boundary value problem arising from the CCH equation. Our approach generalizes the work by Lange [15] to higher order singularly perturbed nonlinear boundary value problems, where standard application of matched asymptotics is not able to locate the position of interior layers that delimit the oscillations of the nonmonotone solutions.

Reyna and Ward [19] previously developed an approach to resolve the internal layer structure of the solutions to the boundary value problem for the related Cahn–Hilliard and viscous Cahn–Hilliard equations. The approach is based on a method due to Ward [25], who uses a "near" solvability condition for the corresponding linearized problem in his asymptotic analysis, and who was inspired by an earlier variational method [13] and work by O'Malley [18] and Rosenblat and Szeto [20], who investigated the problem of spurious solutions to singular perturbation problems of second-order nonlinear boundary value problems [3]. Moreover, for the related Kuramoto–Sivashinsky equation, a multiple-scales analysis of the corresponding third-order nonlinear boundary value problem by Adams, King, and Tew [1] shows that the derivation of monotone and oscillating traveling-wave solutions involve exponentially small terms; their method is based on an analysis of the *Stokes phenomenon* of the corresponding problem in the complex plane (see Howls, Kawai, and Takei [12] for an introduction).

In what follows we begin with the phase-space analysis for the CCH equation in section 2, followed by the asymptotic treatment for $\nu \ll 1$. The asymptotic ideas used for the CCH equation are then applied to the HCCH equation in section 3. The solutions obtained there are useful to serve as initial input for the numerical investigations of the branches of nonmonotone solutions in section 4. In this part we develop our numerical approach and then use it to identify new stationary solutions of the HCCH equation; these agree with the asymptotic theory. Finally, we briefly sum up the results together with concluding remarks in section 5.

2. Stationary solutions of the CCH equation.

The high order term in the CCH equation represents the regularization of the internal layers of the solutions. For most of our investigations we consider the problem in the scaling of the internal layers, or *inner* scaling, where the x-coordinate is stretched around the location $x = \bar{x}$ of a layer according to

$$x^* = \frac{x - \bar{x}}{\varepsilon}. \tag{2.1}$$

In this scaling the CCH equation becomes (after dropping the "$*$")

$$\varepsilon^2 u_t - \frac{\delta}{2}(u^2)_x + (Q(u) + u_{xx})_{xx} = 0, \tag{2.2}$$

where $\delta = \varepsilon\nu$. The stationary problem obtained by setting u_t to zero can be integrated once, requiring that the solutions approach the constants $\pm\sqrt{A}$ as $x \to \mp\infty$, where A is a constant of integration. That is, we consider the boundary value problem

$$\frac{\delta}{2}\left(u^2 - A\right) = (Q(u) + u_{xx})_x \tag{2.3}$$

together with the far-field conditions

$$\lim_{x \to \pm\infty} u = \mp\sqrt{A} \tag{2.4}$$

and vanishing derivatives in the same spatial limit. We refer to solutions of this system as antikinks. Monotone antikinks are known analytically [16], while recently, nonmonotone connections were computed numerically by Zaks et al. [28]. We now briefly discuss the numerical approach for obtaining these solutions. Here we concentrate on the regime where $0 < \delta \ll 1$ in order to compare the results with the asymptotic solutions derived later on. For a bifurcation analysis for larger δ we refer the reader to [28].

2.1. Numerical solutions. For the numerical solutions, we will work with a rescaled version, where we set $u = \sqrt{A}c$ so that the equilibrium points do not depend on A, and for $Q(u)$ given by (1.2), (2.3) and (2.4) become

$$(1 - c^2) = -\frac{2}{\delta\sqrt{A}}(c_{xx} + c - Ac^3)_x\,, \quad \lim_{x\to\pm\infty} c = \mp 1. \tag{2.5}$$

This differential equation could be reduced to second order by the transformation $g(c) = c_x$ at the expense of making it nonautonomous. However, for this problem we find it most convenient to present a shooting method, which enables us to track solution branches in the (A, δ) parameter plane. We transform (2.5) into a first-order system $U' = F(U)$, where $F : \mathrm{R}^3 \to \mathrm{R}^3$ is the function

$$F_1(U) = U_2, \quad F_2(U) = U_3, \quad F_3(U) = (3A(U_1)^2 - 1)U_2 + \frac{\delta\sqrt{A}}{2}((U_1)^2 - 1)\,. \tag{2.6}$$

We work in a three-dimensional phase space and denote either vectors or whole trajectories therein with capital U's. We use the same notation for two different objects because it will be clear from the context what is meant. Subscripts indicate the components.

The characteristic polynomials at the equilibrium points $U^{\pm} = (\pm 1, 0, 0)^T$ are

$$\mathcal{P}^{\pm}(\lambda) = \left|\frac{dF}{dU}(U^{\pm}) - \lambda I\right| = \lambda^3 + \lambda(1 - 3A) \mp \delta\sqrt{A}. \tag{2.7}$$

The signs of the real parts of the roots determine the dimension of the stable and unstable manifolds $W^u(U^+)$, $W^s(U^-)$, $W^s(U^+)$, $W^u(U^-)$ of the equilibrium points. The latter two are two-dimensional and so the existence of a kink is generic, while this is not the case for the antikinks. The dimension of $W^u(U^+)$ and $W^s(U^-)$ is one, so that the heteroclinic connections from the positive to the negative equilibrium arise from a codimension two intersection. This means that with the two parameters A and δ, we can expect only separated solutions when the manifolds intersect, but due to the reversibility properties, which are discussed below, the codimension reduces to one and we can expect separated solutions for the free parameter A and a fixed δ, and hence one or several whole branches in the (A, δ) parameter plane. An example of a nonmonotone connection is sketched in Figure 2.1, where the trajectories wind themselves in the phase space with a solution that exhibits 15 humps.

Reversibility and computations . It is instructive to note that the solution of (2.5) is translation invariant, $c(x) \to c(x + L)$, and that (2.6) forms a reversible dynamical system; hence it is invariant with respect to the transformation $x \to -x, U \to -U$, as has also been noted by Zaks et al. [28].

Let us consider generally a k-dimensional phase space, since the following discussion will be also useful in section 4, where we analyze the HCCH equation with its higher order system. The linear transformation

$$R : \mathrm{R}^k \to \mathrm{R}^k, \quad R(U_j) = (-1)^j U_j, \quad j = 1, \ldots, k, \tag{2.8}$$

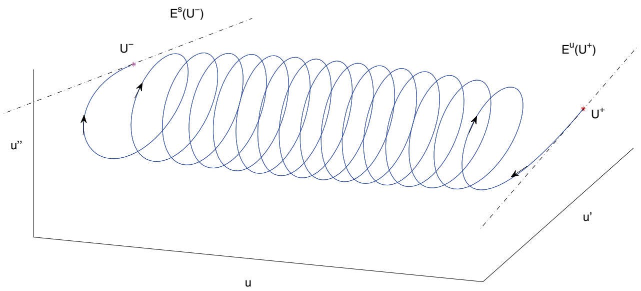


FIG. 2.1. *CCH: Antikink solutions connecting the hyperbolic equilibrium points U^+ and U^- are sought in a three-dimensional phase space. The unstable manifold emerging from U^+, $W^u(U^+)$ is one-dimensional, as is the stable manifold $W^s(U^-)$. The approximating linearized spaces $E^u(U^+)$ and $E^s(U^-)$ are drawn as dash-dotted lines and are used in the computations.*

fulfills $R^2 = Id$ and $RF(U) = -F(RU)$ for $k = 3$ and (2.6) and represents the reversibility in the phase space. It is an involution (or a reflection), and its set of fixed points is the symmetric section of the reversibility; these are zero at odd components, $U_i = 0$ for odd i. A solution that crosses such a point is necessarily symmetric under R, and for each point U on the connection there exists a corresponding transformed point RU somewhere on the branch. In fact there is an equivalence here since odd solutions necessarily cross a point in the symmetric section. It holds that c and its even derivatives have to vanish in the point of symmetry L because of the fulfilled equations $\frac{d^{2m}}{dx^{2m}}c(x+L) = -\frac{d^{2m}}{dx^{2m}}c(-x+L), m = 0, 1, \ldots, \lfloor k/2 \rfloor$, and continuity of the solution and its derivatives.

From the above we conclude that with a shooting method we can stop integrating when we find a point with zero odd components, since the second half of the solution is then given by the set of transformed points under R. Hence we define the following distance function for a trajectory U over the interval of integration which helps to find these points,

$$d_A(U) = \min_x \sqrt{\sum_{i \text{ odd}} U_i(x)^2} \ . \tag{2.9}$$

The minimization of $d_A(U)$ over the free parameter, $\min_A d_A(U)$, must result in the value zero for an antisymmetric heteroclinic solution. We can use this condition for shooting and boundary value problem formulations, for the CCH and later for the HCCH equation in section 4.

For a fixed value of δ and a range of different A we follow the relevant branch of $W^u(U^+)$ by shooting from an initial point $U^+ \pm \epsilon v$ near U^+, where v is a unit eigenvector corresponding to the positive eigenvalue of $dF/dU_{|U=U^+}$ and $\epsilon \ll 1$. We stop the integration if a certain threshold value for $|U_1|$ is crossed. Figure 2.2 shows $d_A(U)$ as a function of A for $\delta = 0.05$.

At this point we have heteroclinic connections for one fixed value of δ, which we denote by $het_k, k = 0, 1, \ldots$ (using the notation in Zaks et al. [28]). het_0 is the analytical, monotone tanh solution, while het_k has k humps on the way down from

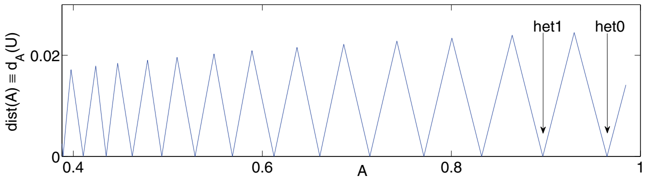


FIG. 2.2. *Distance function d_A defined by* (2.9) *depending on A with fixed $\delta = 0.05$, showing the first* 14 *zeros corresponding to het_0 to het_{13}.*

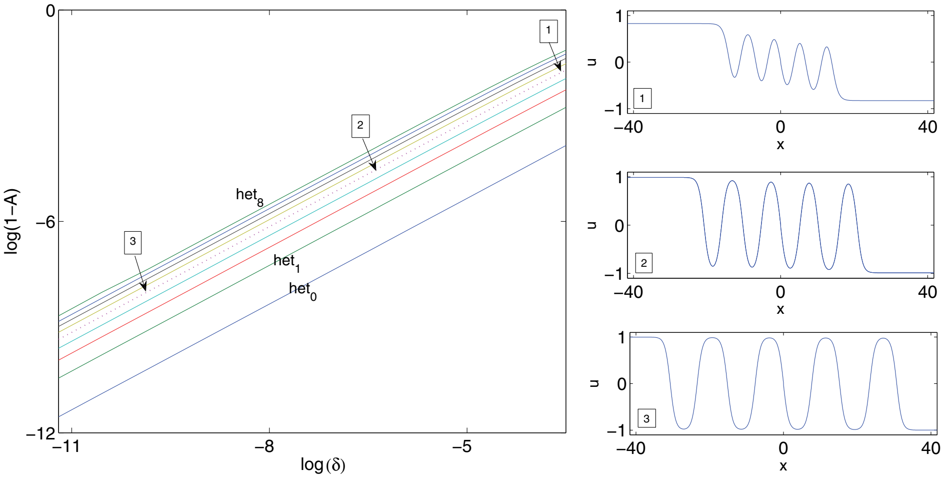


FIG. 2.3. *Parameter plane,* $\log(\delta)$ *for the x-axis and* $\log(1 - A)$ *for the y-axis, for the CCH equation for the first nine antikink solutions $het_k, k = 0, 1, \dots, 8$. The graphs on the right show the shapes of representative het_4 solutions, and hence those on the fifth line from below, for the approximate (A, δ) tuples* $(0.8259, 0.0289)$, $(0.9893, 0.0017)$, *and* $(0.9998, 2.6457 \cdot 10^{-5})$.

$\sqrt{A}$ to $-\sqrt{A}$. We will use the same terminology for the solution structure of the stationary HCCH problem in section 4. Here, a het_k solution corresponds to the kth zero from the right in Figure 2.2. We then follow the roots of the distance function by linearly extrapolating to a new guess for A and use a bisection algorithm to quickly converge to the next root. Figure 2.3 shows a portion of the (A, δ) parameter plane, where we concentrate on very small values of δ or, differently interpreted, on the bifurcation of the various spiraling CCH orbits from the heteroclinic connections of the Cahn–Hilliard equation in its one dimension smaller phase space.

2.2. Asymptotic internal layer analysis. For the asymptotic analysis we use a slightly different scaling than for the numerical treatment. Here, we let

$$x^* = \frac{x - \bar{x}}{\sqrt{2}\,\varepsilon} \tag{2.10}$$

denote the *inner* variable about a layer located at $x = \bar{x}$. For the stationary problem we then obtain

$$\left(u'' + 2\,Q(u)\right)' = \delta\sqrt{2}\left(u^2 - A\right) \tag{2.11}$$

instead of (2.3), where $' = d/dx^*$. For later comparisons of the numerical and asymptotic results we have to keep in mind that the spatial scales differ by a factor of $\sqrt{2}$.

We point out that the problem considered here shares the internal layer structure of the singular perturbation problems discussed by Lange [15], and we will make use of and extend this ansatz for our situation. This will also prove useful for understanding the approach taken for the HCCH problem in section 3.1, since there we have to carefully combine the exponential matching with the conventional matching procedure when matching the two regions. For both problems the asymptotic analysis can be conveniently carried out in terms of the small parameter δ.

In the following analysis we consider the simplest case of a nonmonotone solution with only one hump, as illustrated in Figure 2.4; we note that nonmonotone solutions with more oscillations can be treated similarly.

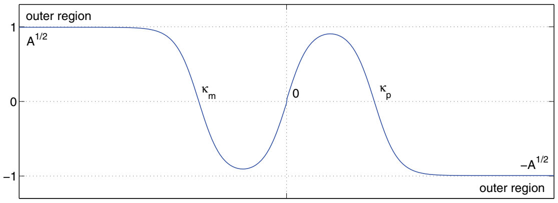


FIG. 2.4. *Sketch of a* 1*-hump, or het*$_1$*, solution showing the general setup for the matching procedure for the CCH and HCCH equations.*

2.2.1. The 1-hump solution. We observe that the 1-hump solution has three internal layers, one at $\kappa_m < 0$, one at κ_p, and one at the symmetry point in between. Since the solution is point symmetric, we can choose this point to be $x = 0$, and it will be enough to discuss only the two layers at κ_m and zero and then match them to the *outer* solution.

Internal layer near κ_m. For the first internal layer at κ_m we let

$$x_m = \frac{x}{\sqrt{2}\,\varepsilon} - \frac{\bar{\kappa}_m}{\sqrt{2}}, \tag{2.12}$$

where $\bar{\kappa}_m < 0$, and set

$$\kappa_m = \bar{\kappa}_m + \sqrt{2}\sum_{k=1}^{\infty} \delta^k \kappa_{mk}, \tag{2.13}$$

so that to leading order the location where the solution crosses zero is $\bar{\kappa}_m$ and the additional terms account for the corrections due to the higher order problems.

With $u_m(x_m) = u(\varepsilon(\bar{\kappa}_m + \sqrt{2}\,x_m))$ the governing equation becomes

$$u_m''' + 2\,Q'(u_m) = \delta\,\sqrt{2}\,(u_m^2 - A)\,, \quad \text{where} \quad ' = \frac{d}{dx_m}. \tag{2.14}$$

For the boundary condition where u_m crosses zero, we have

$$u_m\left(\frac{\kappa_m - \bar{\kappa}_m}{\sqrt{2}}\right) = 0, \tag{2.15}$$

and the condition towards $-\infty$ is

$$\lim_{x_m \to -\infty} u_m(x_m) = \sqrt{A}. \tag{2.16}$$

We now assume $u_m(x_m)$ can be written as the following asymptotic expansion, valid near κ_m:

$$u_\alpha(x_\alpha) = u_{\alpha 0}(x_\alpha) + \sum_{k=1}^{\infty} \delta^k \, u_{\alpha k}(x_\alpha)\,, \tag{2.17}$$

with $\alpha = m$ here. Additionally, we assume A has the asymptotic expansion

$$A = 1 + \delta A_1 + O(\delta^2). \tag{2.18}$$

Observe that from (2.16) and (2.18),

$$\lim_{x_m \to -\infty} u_m(x_m) = \lim_{x_m \to -\infty} u_{m0}(x_m) + \sum_{k=1}^{\infty} \delta^k \, u_{mk}(x_m) = 1 + \frac{1}{2} \sum_{k=1}^{\infty} \delta^k A_k\,. \tag{2.19}$$

To leading order in δ we get the following problem for the Cahn–Hilliard equation:

$$u_{m0}''' + 2\,Q'(u_{m0}) = 0, \tag{2.20a}$$

$$u_{m0}(0) = 0 \quad \text{and} \quad \lim_{x_m \to -\infty} u_{m0}(x_m) = 1 \tag{2.20b}$$

with the unique solution $u_{m0}(x_m) = -\tanh(x_m)$. Next, the problem of order δ is

$$\Big(\mathcal{L}\left(u_{m1}, x_m\right)\Big)' = \sqrt{2}\,\left(\tanh^2(x_m) - 1\right), \tag{2.21a}$$

$$u_{m1}(0) = \kappa_{m1} \quad \text{and} \quad \lim_{x_m \to -\infty} u_{m1}(x_m) = \frac{A_1}{2}\,, \tag{2.21b}$$

where κ_{m1} and A_1 are constants to be exponentially matched and the operator $\mathcal{L}$ is defined by

$$\mathcal{L}(v, z) = v'' + 2\left(1 - 3\,\tanh^2(z)\right)\,v\,, \tag{2.22}$$

and $z = x_m$, $v = u_{m1}$, and $\prime = d/dx_m$. Note that the first boundary condition is obtained by expanding (2.15),

$$\begin{aligned} u_m\left(\sum_{k=1}^{\infty} \delta^k \kappa_{mk}\right) &= u_m\Big(\,\delta\kappa_{m1} + \delta^2\kappa_{m2} + O(\delta^3)\,\Big) \\ &= u_{m0}(0) + \delta\big(\,\kappa_{m1}u_{m0}'(0) + u_{m1}(0)\big) + O(\delta^2), \end{aligned} \tag{2.23}$$

so that collecting the terms of order δ gives

$$u_{m1}(0) = -\kappa_{m1}\,u_{m0}'(0) = \kappa_{m1}\,.$$

Next, we integrate (2.21) once to obtain

$$\mathcal{L}\left(u_{m1}, x_m\right) = f_m(x_m)\,, \tag{2.24}$$

where $f_m(x_m) = -\sqrt{2}\tanh(x_m) + c_m$. Taking the limit of this equation to $-\infty$ yields $c_m = -\sqrt{2} - 2A_1$ so that

$$f_m(x_m) = -\sqrt{2}\,(\tanh(x_m) + 1) - 2A_1\,. \tag{2.25}$$

The homogeneous solutions of (2.24) are

$$\phi_m(x_m) = -u'_{m0}(x_m) = 1 - \tanh^2(x_m)\,, \tag{2.26}$$

$$\psi_m(x_m) = \left(\int_0^{x_m} \frac{dz}{\phi_m^2(z)}\right)\phi_m(x_m)\,. \tag{2.27}$$

Also note that $\lim_{x_m\to-\infty}\phi_m(x_m) = 0$ and $\psi_m(0) = 0$. At this stage it is convenient to choose the inhomogeneous solution that remains bounded as $x_m \to -\infty$ and vanishes at $x_m = 0$, which is satisfied by

$$\varphi_\alpha(x_\alpha) = \psi_\alpha(x_\alpha)\int_{-\infty}^{x_\alpha} \phi_\alpha\, f_\alpha\, dz - \phi_\alpha(x_\alpha)\int_0^{x_\alpha} \psi_\alpha\, f_\alpha\, dz \tag{2.28}$$

with $\alpha = m$. Hence, the unique solution for (2.21) is the linear combination

$$u_{m1}(x_m) = -\kappa_{m1}\phi_m(x_m) + \varphi_m(x_m)\,. \tag{2.29}$$

Internal layer near $x = 0$. For the internal layer near the origin we proceed as above. Here, we stretch the independent variable as

$$x_0 = \frac{x}{\sqrt{2}\varepsilon} \tag{2.30}$$

and construct an asymptotic expansion (2.17) near $x = 0$ with $\alpha = 0$ for the solution of the problem

$$u_0''' + 2\,Q'(u_0) = \delta\,\sqrt{2}\,(u_0^2 - A)\,, \quad \text{where} \quad ' = \frac{d}{dx_0}. \tag{2.31}$$

We note that the point $x = 0$ is assumed to be the symmetry point of the complete solution; hence here we require

$$u_0(0) = 0 \quad \text{and} \quad u_0''(0) = 0\,. \tag{2.32}$$

In anticipation of the exponential matching we also require that $\lim_{x_0\to-\infty} u_{00}(x_0) = -1$, so that the solution to the leading order problem is $u_{00}(x_0) = \tanh(x_0)$. For the solution to $O(\delta)$ we find

$$u_{01}(x_0) = b_0\,\psi_0(x_0) + \varphi_0(x_0), \tag{2.33}$$

where b_0 is a further constant to be exponentially matched. Here, the homogeneous solutions are

$$\phi_0(x_0) = -u'_{00}(x_0) \quad \text{and} \quad \psi_0(x_0) = \left(\int_0^{x_0} \frac{dz}{\phi_0^2(z)}\right)\phi_0(x_0), \tag{2.34}$$

and the inhomogeneous solution is defined by (2.28), where $\alpha = 0$ and $f_0(x_0) = -\sqrt{2}\tanh(x_0)$. They are chosen such that $\varphi_0(0) = 0$ and $\varphi_0''(0) = 0$; in fact we have $\lim_{x_0\to\pm\infty}\varphi_0(x_0) = \pm\sqrt{2}/4$.

2.2.2. Exponential matching. Exponential matching requires that all exponentially small and exponentially growing terms have to be accounted for and matched. This means that first we have to express the variable x_0 in terms of x_m (or vice versa). From the definitions of these variables it follows that

$$x_0 = x_m + \frac{\bar{\kappa}_m}{\sqrt{2}}\,. \tag{2.35}$$

In particular, exponential terms in the solution $u_0(x_0)$ transform as $e^{2x_0} = e^{\sqrt{2}\bar{\kappa}_m}\, e^{2x_m}$ and so forth for higher order exponential terms e^{2nx_0} or terms with different signs in the exponent.

Now note that as $x_0 \to -\infty$, the leading and $O(\delta)$ solutions can be written as

$$u_{00}(x_0) = -1 + 2e^{2x_0} - O(e^{4x_0}), \tag{2.36}$$

and with $\bar{\mu} = \left(\frac{3}{2}b_0 + \sqrt{2}\right)$,

$$u_{01}(x_0) = -\frac{1}{4}\bar{\mu} - \frac{b_0}{16}e^{-2x_0} + \left(\frac{13}{16}b_0 + \frac{1}{\sqrt{2}} + \bar{\mu}x_0\right)e^{2x_0} + O(e^{4x_0})\,. \tag{2.37}$$

Written in x_m variables, the solution

$$\begin{aligned} u_0(x_m) &= -1 + 2e^{2x_m}e^{\sqrt{2}\bar{\kappa}_m} + O(e^{2\sqrt{2}\bar{\kappa}_m}) \\ &\quad + \delta\left(-\frac{1}{4}\bar{\mu} - \frac{b_0}{16}e^{-2x_m}e^{-\sqrt{2}\bar{\kappa}_m} + \left(\frac{13}{16}b_0 + \frac{1}{\sqrt{2}} + \bar{\mu}\left(x_m + \frac{\bar{\kappa}_m}{\sqrt{2}\varepsilon}\right)\right)e^{2x_m}e^{\sqrt{2}\bar{\kappa}_m} \right. \\ &\qquad \left. + O(e^{2\sqrt{2}\bar{\kappa}_m})\right) + O(\delta^2) \end{aligned} \tag{2.38}$$

has to be exponentially matched to

$$\begin{aligned} u_m(x_m) &= -1 + 2e^{-2x_m} + O(e^{-4x_m}) \\ &\quad + \delta\left(-\left(A_1 + \frac{\sqrt{2}}{4}\right) - \frac{1}{4}\left(A_1 + \frac{1}{\sqrt{2}}\right)e^{2x_m} + \left(\frac{7}{2}A_1 + \frac{5}{4}\sqrt{2} + 4\kappa_{m1}\right)e^{-2x_m} \right. \\ &\qquad \left. - \left(3A_1 + \frac{1}{\sqrt{2}}\right)x_m e^{-2x_m} + O\big(e^{-4x_m}\big)\right) + O(\delta^2) \end{aligned} \tag{2.39}$$

as $x_m \to \infty$. While we have already anticipated matching of the constants during the derivation of the leading order solutions, the constant terms of the $O(\delta)$ solutions are first to be matched. Matching to the exponential terms in (2.39) entails a rearranging of terms of different orders of magnitude in the expansion (2.38). In particular, the first exponential term to leading order in (2.39) matches the second term of $O(\delta)$ in (2.38), the second and largest exponential term of $O(\delta)$ in (2.39) matches the second term of the leading order in (2.38), and so forth. Summarizing, we obtain

$$\frac{1}{4}\left(\frac{3}{2}b_0 + \sqrt{2}\right) = A_1 + \frac{\sqrt{2}}{4}, \quad -\rho\frac{b_0}{16} = 2, \quad -\frac{\rho}{4}\left(A_1 + \frac{1}{\sqrt{2}}\right) = 2\,, \tag{2.40}$$

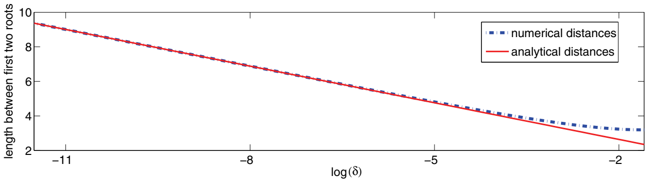


FIG. 2.5. *Distances between the first two roots of the* het_1 *solutions versus* $\log(\delta)$ *together with the width predicted by the asymptotic formula* (2.43).

where we denote $\rho = \delta\, e^{-\sqrt{2}\bar{\kappa}_m}$. Solving yields

$$\rho = 4\sqrt{2}, \quad A_1 = -\frac{3}{\sqrt{2}}, \quad \text{and} \quad b_0 = -\frac{8}{\sqrt{2}}\,. \tag{2.41}$$

We observe that we have determined the $O(\delta)$ correction A_1. Additionally, we now know that $\delta\, e^{-\sqrt{2}\bar{\kappa}_m} = 4\sqrt{2}$, hence

$$\bar{\kappa}_m = \frac{\ln(\delta)}{\sqrt{2}} - \frac{\ln\left(4\sqrt{2}\right)}{\sqrt{2}}, \tag{2.42}$$

and if we recall (2.13) and $\kappa_m < 0$, then the width of the hump is $-\kappa_m$, where

$$\kappa_m = \frac{\ln(\delta)}{\sqrt{2}} - \frac{\ln\left(4\sqrt{2}\right)}{\sqrt{2}} + O(\delta)\ . \tag{2.43}$$

Further constants, such as κ_{m1}, are found by including higher exponential terms and expansions of the higher order problems. Finally, we note that by making use of the symmetry of the solution about the point $x = 0$, the exponential matching of the solution near zero to the one near κ_p proceeds analogously.

2.2.3. Comparison of numerical and asymptotic solution.

For the comparison with the asymptotic solution we are interested mainly in the het_1 solution which we derived in section 2.2.1. By numerical continuation of the shooting method, one obtains N tuples $(A^{(j)}, \delta^{(j)})$, $j = 1, \ldots, N$, in the parameter plane that give a het_1-branch when being connected. We use two vectors of parameters which we abbreviate $\mathbf{A} = (A^{(j)})_{j=1,\ldots,N}$ and $\boldsymbol{\delta} = (\delta^{(j)})_{j=1,\ldots,N}$ to confirm the formulas we obtained in the previous section. Further we make use of a distance vector $\mathbf{K} = (K^{(j)})_{j=1,\ldots,N}$. $\mathbf{K}$ contains the distances between the zero crossings of the solutions, or in context of the asymptotics section (see Figure 2.4) $K^{(j)} \approx |\kappa_m(\delta^{(j)})|$. To obtain the relation between A and δ and the evolution of the distances we solve the least squares problems

$$\min_{\mu_1} \|(\mathbf{1} - \mu_1\boldsymbol{\delta}) - \mathbf{A}\|_2^2 \quad \text{and} \quad \min_{\eta_1,\eta_2} \|\eta_1 \log(\boldsymbol{\delta}\eta_2) - \mathbf{K}\|_2^2,$$

and hence we assume a linear law for the A-values in δ and a general logarithmic law for the distances. We obtain

$$A \approx 1 - 2.12\delta \approx 1 - \frac{3}{\sqrt{2}}\delta \quad \text{and} \quad K \approx -0.71\log(0.18\delta), \tag{2.44}$$

which confirms the results from the analysis, (2.41) and (2.43). We see the good match in the distance plot in Figure 2.5.

These results motivated us to obtain a general rule for the relation between the two parameters of the CCH equation for different stationary solutions. The numerically computed branches in Figure 2.3 show that the slopes of the het_k branches are one when plotting $\log(\delta)$ against $\log(1-A)$, so that the relation $\log(\delta)+\text{const} = \log(1-A)$ shows the linear dependence $A(k) = 1+A_1(k)\delta$, where $A(k)$ is the A value for the het_k solution and $A_1(k)$ its linear coefficient. We see that the magnitude of A_1 increases linearly with the order k of the heteroclinic connection, and we obtain a general expression for the squared far-field value for nonmonotone het_k solutions, namely,

$$A_1(k) = -\frac{2k+1}{\sqrt{2}}\,. \tag{2.45}$$

3. Matched and exponential asymptotics for the stationary HCCH equation. As we did for the CCH equation, we will perform our analysis of the internal layers in the inner scaling (2.10). From the stationary form of (1.4) we obtain the equation

$$\left(u'' + 2\,Q(u)\right)''' = -\delta\; 2^{3/2}\left(u^2 - A\right) \tag{3.1}$$

after integrating once and requiring that for an antikink $\lim_{x\to\pm\infty} u = \mp\sqrt{A}$ and setting $\delta = \epsilon^3\nu$ here. We consider the het_1 (1-hump) solution and again make use of the point symmetry of the problem. Now, however, unlike for the CCH equation, the solutions in the *outer* region are not just constants. Here, we have to introduce an *outer* layer to the left of the *inner* layer about κ_m; see also Figure 2.4 for the case of a 1-hump solution. In the following subsections we first briefly derive the solution to this *outer* problem and match it to the solution to the *inner* problem near κ_m. The remaining degrees of freedom are then used to exponentially match it to a second inner layer near $x = 0$.

It has been demonstrated in [22] for monotone antikink solutions of the HCCH equation that it is necessary to match terms up to order δ in order to obtain the correction A_1, given the asymptotic expansion of A,

$$A = 1 + \sum_{k=1}^{\infty} \delta^{k/3}\, A_k\,. \tag{3.2}$$

Here, for the nonmonotone antikinks we have to match *inner* and *outer* solutions and then also exponentially match the *inner* layers. This has to be carried through iteratively up to three orders of magnitude in order to obtain not only the correction A_1 but also the expression for the width of the humps.

3.1. The 1-hump solution for the HCCH equation. We start by shifting to the inner coordinates that describe the region near κ_m, which is to be matched to the outer region. Again defining x_m by (2.12), the governing equation in this inner region is

$$\Big(u_m'' + 2\,Q(u_m)\Big)''' = -\,2^{3/2}\,\delta\,(u_m^2 - A), \quad \text{where} \quad {}' = \frac{d}{dx_m}. \tag{3.3}$$

For the boundary conditions we again place κ_m near the point where u_m crosses zero, i.e.,

$$u_m\left(\frac{\kappa_m - \bar{\kappa}_m}{\sqrt{2}}\right) = 0\,. \tag{3.4}$$

The condition towards $-\infty$ is not as trivial as for the CCH equation but needs to be matched to the outer solution in the region to the left of κ_m (or to the right of κ_p, taking account of symmetry).

For the *outer* region (see Figure 2.4), where x_m becomes very large, we use the ansatz

$$\xi = \delta^{1/3}\, x_m \quad \text{and} \quad Y(\xi;\delta) = u_m(x_m;\delta) \tag{3.5}$$

and obtain the *outer* problem

$$\Big(\delta^{2/3}\, Y_{\xi\xi} + 2\, Q\,(Y)\Big)_{\xi\xi\xi} = -\,2^{3/2}\,\left(Y^2 - A\right) \tag{3.6}$$

with the far-field condition

$$\lim_{\xi\to-\infty} Y(\xi) = \sqrt{A}\,. \tag{3.7}$$

The region near $x = 0$, for which we use the variable x_0 from (2.30), is described by the problem

$$\Big(u_0'' + 2\, Q(u_0)\Big)''' = -\,2^{3/2}\,\delta\,(u_0^2 - A) \quad \text{where} \quad {}' = \frac{d}{dx_0}. \tag{3.8}$$

The point $x = 0$ is the point of symmetry of the solution. Here we require

$$u_0(0) = 0, \quad u_0''(0) = 0, \quad \text{and} \quad u_0''''(0) = 0\,, \tag{3.9}$$

plus additional conditions from the exponential matching to the internal layer near κ_m as $x_0 \to -\infty$, as we have shown for the CCH equation.

Here we assume that the solutions to these three problems for Y, u_m, and u_0 can be represented by asymptotic expansions

$$u_\alpha(x_\alpha;\varepsilon) = u_{\alpha 0}(x_\alpha) + \sum_{k=1}^{\infty} \delta^{k/3}\, u_{\alpha k}(x_\alpha), \quad \text{where} \quad \alpha = 0, m, \tag{3.10}$$

valid near κ_m and $x = 0$, respectively, and

$$Y(\xi;\delta) = Y_0(\xi) + \sum_{k=1}^{\infty} \delta^{k/3}\, Y_k(\xi)\,, \tag{3.11}$$

valid in the outer region, where we let

$$\kappa_m = \bar{\kappa}_m + \sqrt{2}\sum_{k=1}^{\infty} \delta^{k/3}\kappa_{mk}\,. \tag{3.12}$$

Obtaining solutions to the outer problem is straightforward [22], but in order to be more comprehensible we include the results in Appendix A. The solutions to the other regions are discussed now.

3.1.1. Leading order. To leading order in δ we get the problem

$$(u''_{m0} + 2\,Q(u_{m0}))''' = 0, \tag{3.13a}$$

$$u_{m0}(0) = 0. \tag{3.13b}$$

Matching to the leading order outer solution (A.2) $Y_0 = 1$ we find

$$u_{m0}(x_m) = -\tanh(x_m)\,. \tag{3.14}$$

Its representation towards the internal layer about $x = 0$ is given by

$$u_{m0} = -1 + 2e^{-2x_m} - 2e^{-4x_m} + O(e^{-6x_m}) \tag{3.15}$$

as $x_m \to \infty$. The leading order problem for this region is

$$(u''_{00} + 2\,Q(u_{00}))''' = 0, \tag{3.16a}$$

$$u_{00}(0) = 0, \quad u''_{00}(0) = 0, \quad \text{and} \quad u''''_{00}(0) = 0, \tag{3.16b}$$

and its solution is

$$u_{00}(x_0) = \tanh(x_0)\,. \tag{3.17}$$

As $x_0 \to -\infty$, its behavior is given by

$$u_{00} = -1 + 2e^{2x_0} - 2e^{4x_0} + O(e^{6x_0})\,. \tag{3.18}$$

3.1.2. $\mathbf{O(\delta^{1/3})}$.

Internal layer near $x = \kappa_m$. The expansion of (3.3) and (3.4) to order $\delta^{1/3}$ yields

$$\mathcal{L}(u_{m1}, x_m) = f_{m1}(x_m), \tag{3.19a}$$

$$u_{m1}(0) = -u'_{m0}(0)\,\kappa_{m1} = \kappa_{m1}, \tag{3.19b}$$

where $\mathcal{L}$ is defined by (2.22) as for the CCH equation and

$$f_{m1}(x_m) := c_{1m}x_m^2 + c_{2m}x_m + c_{3m}\,. \tag{3.20}$$

The homogenous solutions are therefore (2.26) and (2.27). The constants c_{1m}, c_{2m}, c_{3m} are obtained by three successive integrations of the ODE for u_{m1} obtained at this order. We choose the inhomogeneous solution so that it grows only algebraically as $x_m \to -\infty$ and vanishes at $x_m = 0$. Particular solutions to (3.19b) are of the form

$$\varphi_{\alpha j}(x_\alpha) = \psi_\alpha(x_\alpha) \int_0^{x_\alpha} \phi_\alpha\, f_{\alpha j}\, dz - \phi_\alpha(x_\alpha) \int_0^{x_\alpha} \psi_\alpha\, f_{\alpha j}\, dz + \gamma_{\alpha j}\psi_\alpha(x_\alpha)\,, \tag{3.21}$$

so that now we obtain φ_{m1} for $\alpha = m, j = 1$ in (3.21) and

$$\gamma_{m1} = -\frac{\pi^2}{12}c_{1m} + \ln(2)c_{2m} - c_{3m}\,. \tag{3.22}$$

Hence the solution is

$$u_{m1}(x_m) = -\kappa_{m1}\phi_m(x_m) + \varphi_{m1}(x_m)\,. \tag{3.23}$$

We evaluate ψ_α, ϕ_α, etc. and subsequent functions with the assistance of *Maple*. As $x_m \to -\infty$, the limiting behavior of u_{m1} is

(3.24)
$$\begin{aligned} u_{m1}(x_m) =& -\frac{1}{8}(c_{1m}+2c_{3m}) - \frac{1}{4}c_{2m}x_m - \frac{1}{4}c_{1m}x_m^2 \\ &+ \left(\frac{1}{64}(-7c_{1m} - 8c_{3m} + 256\kappa_{m1} + 30c_{2m} + 4c_{2m}\pi^2 - 72c_{1m}\zeta(3))\right. \\ &\left.+\frac{1}{16}(-6c_{2m}+15c_{1m}+24c_{3m})x_m + \frac{1}{8}(6c_{2m}-3c_{1m})x_m^2 + \frac{1}{2}c_{1m}x_m^3\right) e^{2x_m} \\ &+O(\,e^{4x_m}), \end{aligned}$$

where ζ is the Riemann zeta function, and u_{m1} must match the outer solution, which is given in the appendix by (A.10) and has only constant terms to this order. Hence we require $c_{2m} = 0$ and $c_{1m} = 0$. The matched solution is now

$$\begin{aligned} u_{m1}^{(m)}(x_m) &= (1-\tanh^2(x_m))\,\kappa_{m1} \\ &\quad -\frac{c_{3m}}{16}\Big(-2e^{6x_m} + 4 + 10e^{2x_m} - 12e^{4x_m} - 24x_m e^{2x_m}\Big)\frac{e^{-2x_m}}{(e^{2x_m}+1)^2}\,, \end{aligned} \tag{3.25}$$

where we denote by $u_{m1}^{(m)}$ the solution that is obtained by matching to the outer solution Y. As we will see later, exponential matching to the inner solution u_0, i.e., as $x_m \to \infty$, where we find

$$\begin{aligned} u_{m1}^{(m)}(x_m) = \frac{1}{8}c_{3m}e^{2x_m} + \frac{1}{2}c_{3m} + \left(-\frac{7}{4}c_{3m} + 4\kappa_{m1} + \frac{3}{2}c_{3m}x_m\right)e^{-2x_m} \\ + \left(\frac{11}{4}c_{3m} - 8\kappa_{m1} - 3c_{3m}x_m\right)e^{-4x_m} + O(e^{-6x_m})\,, \end{aligned}$$

requires also $c_{3m} = 0$. Hence, denoting by $u_{m1}^{(e)}$ the solution that has been exponentially matched to the inner solution u_0 near $x = 0$, we obtain

$$u_{m1}^{(e)}(x_m) = \big(1-\tanh^2(x_m)\big)\;\kappa_{m1}\,. \tag{3.26}$$

Internal layer near $x = 0$. The $O(\delta^{1/3})$ problem is

$$\mathcal{L}(u_{01}, x_0) = f_{01}(x_0)\,, \tag{3.27a}$$
$$u_{01}(0) = 0, \quad u_{01}''(0) = 0, \quad \text{and} \quad u_{01}''''(0) = 0\,, \tag{3.27b}$$

with

$$f_{01}(x_0) := c_{10}x_0^2 + c_{20}x_0 + c_{30}\,. \tag{3.28}$$

Its general solution reads

$$u_{01}(x_0) = \varphi_{01}(x_0) + g_1\,\psi_0(x_0)\,, \tag{3.29}$$

where the homogeneous solutions are as before and the inhomogeneous solution is given by (3.21) with $\alpha = 0, j = 1$, and

$$\gamma_0 = -\frac{\pi^2}{12}c_{10} + \ln(2)\,c_{20} - c_{30}\,, \tag{3.30}$$

so that $\varphi_{01}(0)=0$ and φ_{01} grows algebraically as $x_0 \to -\infty$. Furthermore, symmetry requires $\varphi''_{01}(0)=0$ and $\varphi''''_{01}(0)=0$, which implies $c_{10}=0$ and $c_{30}=0$, leading to

$$\begin{aligned}\varphi_{01}(x_0) = {}& \frac{c_{20}}{16(1+e^{-2x_0})^2}\Big(1-4x_0+12\,\mathrm{dilog}(e^{2x_0}+1)e^{-2x_0}-e^{-4x_0}+12x_0^2e^{-2x_0}\\ &+\pi^2e^{-2x_0}+12x_0e^{-4x_0}-14x_0e^{-2x_0}-\ln(1+e^{-2x_0})e^{2x_0}+8e^{-4x_0}\ln(1+e^{-2x_0})\\ &-8\ln(1+e^{-2x_0})+e^{-6x_0}\ln(1+e^{-2x_0})+2e^{-6x_0}x_0\Big)\,,\end{aligned} \tag{3.31}$$

where dilog denotes the dilogarithm function. The remaining free parameters of u_{01} to be matched are c_{20} and g_1. As will be demonstrated later, exponential matching to u_m requires an expression for u_{01} as $x_0 \to -\infty$,

$$\begin{aligned}u_{01}(x_0) = {}& -\frac{g_1}{16}e^{-2x_0}-\frac{1}{4}c_{20}x_0-\frac{3}{8}g_1\\ &+\frac{1}{32}\Big(2c_{20}\pi^2+15c_{20}+26g_1+(48g_1-12c_{20})x_0+24c_{20}x_0^2\Big)e^{2x_0}\\ &+\frac{1}{48}\Big(-36g_1-89c_{20}-6c_{20}\pi^2+(84c_{20}-144g_1)x_0-72c_{20}x_0^2\Big)e^{4x_0}+O(e^{6x_0}),\end{aligned} \tag{3.32}$$

and then re-expanding u_0 in the variable x_m. This shows that also $c_{20}=0$, $g_1=0$ and $c_{3m}=0$. Any other choice leads to a system for the parameters having no solution. Hence, only κ_m remains as a free constant in the two regions. The exponentially matched solution is therefore simply

$$u_{01}^{(e)}(x_0)=0\,. \tag{3.33}$$

3.1.3. O($\delta^{2/3}$).

Internal layer near κ_m. The problem of order $\delta^{2/3}$ is

$$\mathcal{L}(u_{m2},x_m)=f_{m2}(x_m)\,, \tag{3.34a}$$

$$u_{m2}(0)=-u'_{m0}(0)\,\kappa_{m2}-\frac{1}{2}u''_{m0}\kappa_{m1}^2-u'_{m1}(0)\kappa_{m1}=\kappa_{m2}-u'_{m1}(0)\,\kappa_{m1}\,, \tag{3.34b}$$

where

$$f_{m2}(x_m):=d_{1m}x_m^2+d_{2m}x_m+d_{3m}+6\,u_{m0}\,(u_{m1}^{(e)})^2\,. \tag{3.35}$$

Note that $u_{m1}^{(m)'}(0)=0$. Again we choose the inhomogeneous solution so that it grows only algebraically as $x_m \to -\infty$ and vanishes at $x_m=0$ to obtain (3.21) with $\alpha=m, j=2$, and

$$\gamma_{m2}=-\frac{\pi^2}{12}d_{1m}+\ln(2)\,d_{2m}-d_{3m}-\kappa_{m1}^2\,, \tag{3.36}$$

so that the general solution is represented as

$$u_{m2}(x_m)=-\kappa_{m2}\phi_m(x_m)+\varphi_{m2}(x_m)\,. \tag{3.37}$$

As $x_m \to -\infty$, we have to compare

$$\begin{aligned} u_{m2}(x_m) = & -\frac{1}{8}(d_{1m} + 2d_{3m}) - \frac{1}{4}d_{2m}x_m - \frac{1}{4}d_{1m}x_m^2 \\ & + e^{2x_m}\Bigg(\frac{1}{64}[(-7 - 72\zeta(3))d_{1m} - 8d_{3m} + 256(\kappa_{m2} - \kappa_{m1}^2) + (30 + 4\pi^2)d_{2m}] \\ & + \frac{3}{16}(5d_{1m} - 2d_{2m} + 8d_{3m})x_m + \frac{3}{8}(2d_{2m} - d_{1m})x_m^2 + \frac{1}{2}d_{1m}x_m^3\Bigg) + O(e^{4x_m}) \end{aligned}$$

with the outer solution. Matching the constant and the linear terms in x_m yields

$$-\frac{1}{4}d_{3m} = \frac{1}{2}A_1 - \frac{1}{8}A_1^2 + \frac{1}{3}C_1A_1 + \frac{23}{14}C_1^2 + D_1\,, \tag{3.38}$$

$$-\frac{1}{4}d_{2m} = 2^{1/6}C_1\,. \tag{3.39}$$

There is no quadratic term in the outer solution (A.10); hence $d_{1m} = 0$. There are further matching conditions, but they do not simplify the problem structurally at this point and will be enforced later, so that d_{2m}, d_{3m}, and κ_{m2} remain to be determined via exponential matching. As $x_m \to \infty$, the expansion to this order can be written as

(3.40)

$$\begin{aligned} u_{m2}^{(m)} = & \frac{1}{2}d_{3m} - \frac{1}{4}d_{2m}x_m + \frac{1}{8}d_{3m}e^{2x_m} + \frac{e^{-2x_m}}{32}\Big(- 56d_{3m} - 15d_{2m} \\ & - 2d_{2m}\pi^2 + 128(\kappa_{m1}^2 + \kappa_{m2}) + (48d_{3m} - 12d_{2m})x_m - 24d_{2m}x_m^2\Big) + O(e^{-4x_m})\,. \end{aligned}$$

Internal layer near $x = 0$. As for the $O(\delta^{1/3})$ problem, at $O(\delta^{2/3})$ we have

$$\mathcal{L}(u_{02}, x_0) = f_{02}(x_0)\,, \tag{3.41a}$$

$$u_{02}(0) = 0, \quad u_{02}''(0) = 0, \quad \text{and} \quad u_{02}''''(0) = 0\,, \tag{3.41b}$$

with

$$f_{02}(x_0) := d_{10}x_0^2 + d_{20}x_0 + d_{30}\,. \tag{3.42}$$

The general solution is

$$u_{02}(x_0) = \varphi_{02}(x_0) + g_2\,\psi_0(x_0)\,, \tag{3.43}$$

where the homogeneous component is as before and the inhomogeneous part is obtained by setting $\alpha = 0$, $j = 2$, and $\gamma_{02} = 0$ in (3.21), so that $\varphi_{02}(0) = 0$ and φ_{02} grows algebraically as $x_0 \to -\infty$. Symmetry requires $\varphi_{02}''(0) = 0$, $\varphi_{02}''''(0) = 0$, which implies $d_{10} = 0$ and $d_{30} = 0$. The remaining free parameters to be matched are d_{20} and g_2. In order to exponentially match to u_m to $O(\delta^{2/3})$ and obtain $u_{m2}^{(e)}$, we again have to expand $u_{02}(x_0)$ as $x_0 \to -\infty$, giving

(3.44)

$$\begin{aligned} u_{02}(x_0) = & -\frac{\hat{\mu}}{16}e^{-2x_0} - \frac{1}{4}d_{20}x_0 - \frac{3}{8}\hat{\mu} \\ & + \frac{1}{32}\Big((15 + 2\pi^2 + 2\ln(2))d_{20} + 26g_2 + (48\hat{\mu} - 12d_{20})x_0 + 24d_{20}x_0^2\Big)e^{2x_0} \\ & + \frac{1}{48}\Big(- (89 + 6\pi^2)d_{20} - 36\hat{\mu} + (84d_{20} - 144\hat{\mu})x_0 - 72d_{20}x_0^2\Big)e^{4x_0} + O(e^{6x_0})\,, \end{aligned}$$

and re-express in terms of x_m, where we have used the abbreviation $\hat{\mu} = d_{20}\ln(2) + g_2$.

3.1.4. O(δ).

Internal layer near κ_m. The problem to be solved at order $O(\delta)$ is

$$\mathcal{L}(u_{m3}, x_m) = f_{m3}(x_m)\,, \tag{3.45a}$$

$$\begin{aligned} u_{m3}(0) = {} & -u'_{m2}(0)\kappa_{m1} - u''_{m0}(0)\kappa_{m1}\kappa_{m2} - u'_{m0}(0)\kappa_{m3} \\ & - \frac{1}{6}u'''_{m0}(0)\kappa_{m1}^3 - u'_{m1}(0)\kappa_{m2} - \frac{1}{2}u''_{m1}(0)\kappa_{m2}^2\,, \end{aligned} \tag{3.45b}$$

with

$$\begin{aligned} f_{m3}(x_m) := {} & 2\left((u_{m1}^{(e)})^3 + 6\,u_{m0}\,u_{m1}^{(e)}\,u_{m2}^{(e)}\right) \\ & -2^{3/2}\left[\frac{1}{2}\mathrm{dilog}(e^{2x_m}+1) + \frac{1}{2}(1+k_{1m})x_m^2 + (\ln(2)+k_{2m})x_m + k_{3m}\right]. \end{aligned} \tag{3.46}$$

Again we choose the inhomogeneous solution so that it grows only algebraically as $x_m \to -\infty$ and vanishes at $x_m = 0$ and so that we obtain $\varphi_{m3}(x_m)$ by using formula (3.21) with $\alpha = m$, $j = 3$, and $\gamma_{m3} = 0$. The solution is

$$u_{m3}(x_m) = -u_{m3}(0)\phi_m(x_m) + \varphi_{m3}(x_m)\,, \tag{3.47}$$

where k_{1m}, k_{2m}, k_{3m}, and κ_{m3} remain to be determined via matching. In order to exclude exponential growth as $x_m \to -\infty$ we obtain the relation

$$\begin{aligned} k_{2m} = \frac{\sqrt{2}}{48\ln(2)}\Big(& \kappa_{m1}\left(-(12+9\pi^2)d_{2m} + 12d_{3m} - 24\kappa_{m2}\right). \\ & +\sqrt{2}(24k_{3m} - 12\ln(2)^2 + k_{1m}\pi^2)\Big)\,, \end{aligned} \tag{3.48}$$

so that the expansion obtained as $x_m \to -\infty$ is

$$\begin{aligned} u_{m3}(x_m) = {} & \frac{1}{4\sqrt{2}}(1+k_{1m}+4k_{3m}) + \frac{1}{\sqrt{2}}(\ln(2)+k_{2m})x_m \\ & + (k_{1m}+1)\frac{\sqrt{2}}{4}x_m^2 + O(e^{2x_m})\,. \end{aligned} \tag{3.49}$$

Comparing this with the outer solution to $O(\delta)$, equation (A.10), yields the matching conditions

$$\begin{aligned} \frac{1}{4\sqrt{2}}(1+k_{1m}+4k_{3m}) = {} & \left(-\frac{1}{4}A_1 + \frac{1}{3}C_1\right)A_2 + \left(\frac{7}{12}C_1^2 + \frac{1}{3}D_1\right)A_1 \\ & +\frac{1}{2}A_3 - \frac{59}{216}C_1A_1^2 - \frac{1}{12}2^{1/3}C_1 + K_1 - \frac{23}{7}C_1D_1 + \frac{1}{16}A_1^3 + \frac{127}{28}C_1^3 \end{aligned} \tag{3.50}$$

for the constant terms,

$$\frac{1}{\sqrt{2}}(\ln(2)+k_{2m}) = \left(D_1 - \frac{23}{7}C_1^2\right)2^{1/6} \quad \text{and} \quad (k_{1m}+1)\frac{\sqrt{2}}{4} = 2^{-2/3}C_1 \tag{3.51}$$

for the linear and the quadratic terms, respectively.

Expanding the solution as $x_m \to \infty$, we find

$$
\begin{aligned}
u_{m3}(x_m) = {} & \frac{1}{192}\Big(\kappa_{m1} d_{2m}(9\pi^2+24) - 48\kappa_{m1} d_{3m} + 2\sqrt{2}\pi^2(1-k_{1m}) - 48\sqrt{2}k_{3m}\Big)e^{2x_m} \\
& + \frac{1}{96}\left(\kappa_{m1} d_{2m}(27\pi^2+72) + \sqrt{2}(k_{1m}(12-6\pi^2) - 96k_{3m} - 12 + 2\pi^2)\right) \\
& + \frac{1}{\sqrt{2}}(\ln(2)+k_{2m})x_m + (k_{1m}+1)\frac{\sqrt{2}}{4}x_m^2 + O(e^{-2x_m})\,,
\end{aligned}
\tag{3.52}
$$

and we will exponentially match it to the solution near $x = 0$, which we construct next.

Internal layer near $x = 0$. The general solution to the $O(\delta)$ problem

$$\mathcal{L}(u_{03}, x_0) = f_{03}(x_0)\,, \tag{3.53a}$$

$$u_{03}(0) = 0, \quad u_{03}''(0) = 0, \quad \text{and} \quad u_{03}''''(0) = 0\,, \tag{3.53b}$$

with

$$f_{03}(x_0) := -2^{1/2}\left[\operatorname{dilog}(e^{2x_0}+1) - \operatorname{dilog}(2) + 2\mu_2 x_0 + (1+k_{10})x_0^2\right] \tag{3.54}$$

and the abbreviation $\mu_2 = \ln(2) + k_{20}$, reads

$$u_{03}(x_0) = \varphi_{03}(x_0) + g_3\,\psi_0(x_0)\,, \tag{3.55}$$

where we have required that $u_{03}(0) = 0$ and $u_{03}''(0) = 0$. If we also enforce $u_{03}''''(0) = 0$, then $k_{10} = 0$. Again we take an inhomogeneous solution $\varphi_{03}(x_0)$, which satisfies the above conditions, so that the general solution is obtained with

$$\mu_1 = \sqrt{2}(\ln(2)^2 + 2k_{20}\ln(2)) - g_3 \quad \text{and} \quad \omega = \int_0^1 \frac{1}{z}\ln\left(\frac{z^2+1}{2z}\right)^2 - \frac{\ln(2z)^2}{z}dz \approx 0.3094\,,$$

as

$$
\begin{aligned}
u_{03} = {} & \frac{12\mu_1 - \pi^2\sqrt{2}}{192}e^{-2x_0} + \frac{1}{96}(36\mu_1 + \sqrt{2}(12-\pi^2)) + \frac{\mu_2}{\sqrt{2}}x_0 + \frac{\sqrt{2}}{4}x_0^2 \\
& + \left[\frac{1}{192}\left(156\mu_1 + \sqrt{2}[(19-24k_{20})\pi^2 - 15 - 288\omega - 180\mu_2]\right)\right. \\
& \left. + \frac{1}{16}\Big(-24\mu_1 + \sqrt{2}(12\mu_2 - 11)\Big)\,x_0 + \frac{\sqrt{2}}{8}\,(3 - 12\mu_2)\,x_0^2 - \frac{1}{\sqrt{2}}x_0^3\right]e^{2x_0} + O(e^{4x_0})\,.
\end{aligned}
\tag{3.56}
$$

For exponentially matching to u_m this again has to be re-expressed in x_m and combined with the corresponding expressions for u_{00}, u_{01}, and u_{02} . This will be done in the next section.

3.2. Exponential matching. Now we have to match the rest of the solution $u_m(x_m)$ to the rest of the solution $u_0(x_0)$. This requires matching the exponential terms in addition to the algebraic terms, similarly to the procedure for the CCH equation; i.e., matching of the solution describing the internal layer near $x = \kappa_m$ to the solution near $x = 0$ requires expressing the variable x_0 in terms of x_m (or vice versa). Recall again that $x_0 = x_m + \bar{\kappa}_m/\sqrt{2}$ and that $\bar{\kappa}_m < 0$; the e^{2x_0} terms in the u_0 expansion will produce e^{2x_m} terms with a factor $e^{\sqrt{2}\bar{\kappa}_m}$ (and analogously for e^{-2x_0} terms) and so we will find their corresponding matching partner at a different order in δ in the u_m expansion, as we have shown for the CCH equation. The somewhat subtle difference here is that additionally we need to determine the relationship between $e^{\sqrt{2}\bar{\kappa}_m}$ and δ, and we have in principle several choices, only one of which allows a consistent matching of both expansions. One can observe that the choice $e^{\sqrt{2}\bar{\kappa}_m} = \rho\,\delta^{1/3}$, where ρ is some constant, quickly leads to a contradiction. However, setting

$$e^{\sqrt{2}\bar{\kappa}_m} = \rho\,\delta^{2/3} \tag{3.57}$$

will lead to an $O(\delta^{2/3})$ shift of terms, so that, e.g.,

$$e^{2x_0} \quad \text{will shift to a term} \quad \delta^{2/3}\,e^{2x_m}\,, \tag{3.58}$$

$$e^{-2x_0} \quad \text{will end up as a term} \quad \delta^{-2/3}\,e^{-2x_m}\,, \tag{3.59}$$

and so forth, so that, e.g., a term e^{2x_0} in the leading order part of the u_0 expansion will have to match an e^{2x_m} term in the $O(\delta^{2/3})$ part of the u_m expansion, or an e^{-2x_0} term in the $O(\delta)$ part of the u_0 expansion will have to match an e^{-2x_m} term in the $O(\delta^{1/3})$ part of the u_m expansion. This will also produce terms that will have no partner term in the transformed expansion. Their coefficients must then be set to zero. If we now sum the expansions for $u_{01}(x_0)$, $u_{02}(x_0)$, and $u_{03}(x_0)$ and re-expand using (3.57), we obtain

$$\begin{aligned} u_0(x_m) = {}& -1 - \frac{1}{16}\Big(d_{20}\ln(2) + g_2\Big)e^{-2x_m}\rho + \frac{1}{192}\Big(12\mu_1 - \sqrt{2}\pi^2\Big)e^{-2x_m}\,\rho\,\delta^{1/3} \\ & + \frac{1}{24}\Big(d_{20}(3\ln(\rho) - 9\ln(2) - 2\ln(\delta)) - 9g_2 - 6d_{20}x_m + 48e^{2x_m}/\rho\Big)\delta^{2/3} \\ & + \Bigg[\frac{1}{96}\Bigg(36\mu_1 + \sqrt{2}\Big[12 + (16\ln(\delta) - 24\ln(\rho))\mu_2 \\ & \qquad + 6\left(\ln(\rho) - \frac{2}{3}\ln(\delta)\right)^2 - \pi^2\Big]\Bigg) \\ & + \frac{\sqrt{2}}{12}\,(2\ln(\delta) + 6\mu_2 - 3\ln(\rho))\,x_m + \frac{\sqrt{2}}{4}x_m^2\Bigg]\delta\,, \end{aligned} \tag{3.60}$$

which has to match $u_{m1}(x_m)$, $u_{m2}(x_m)$, and $u_{m3}(x_m)$ to each order, respectively. From this we obtain further conditions for the parameters in addition to those we have already found. Solving the complete system of equations then yields the solutions for the width of the hump

$$\Delta = \frac{\sqrt{2}}{6}\ln\left(\frac{\beta}{W(\beta^{1/3})^3}\right)\,, \tag{3.61}$$

with $\beta = 2^{11}/(27\delta^2)$, where W is the Lambert W function (so $W(x)$ is the solution of $x = W\exp(W)$). The expressions for the remaining matching constants C_1, D_1, etc. are omitted. The first correction in (3.2) has the coefficient

$$A_1 = -3\,2^{1/6}\,. \tag{3.62}$$

Note that not only the transformed expansions but also the expressions for the parameters contain so-called *logarithmic switch-back* terms.

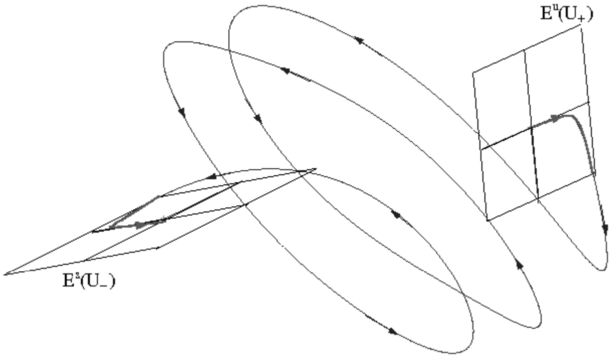


FIG. 4.1. *HCCH: Heteroclinic orbits between the equilibrium points are sought in a five-dimensional phase space that is indicated here in three dimensions. The manifolds $W^u(U^+)$ and $W^s(U^-)$ are two-dimensional, which is suggested by the two planes in the picture.*

4. Numerical method for the fifth-order phase space.

For the numerical stationary solutions of the HCCH equation (3.1) we apply the same scaling for u that we used for the CCH equation to obtain equilibrium points at ±1,

$$(1-c^2)=\frac{2}{\delta\sqrt{A}}(c_{xx}+c-Ac^3)_{xxx}\,,\qquad \lim_{x\to\pm\infty} c=\mp 1, \tag{4.1}$$

again assuming that derivatives vanish in the far field. Reduction to a first-order system $U'=F(U)$, with $F:\mathrm{R}^5\to\mathrm{R}^5$, gives a five-dimensional phase space, where the first four components of $F_i(U)$ are equal to U_{i+1} and the fifth is

$$F_5(U)=6A(U_2)^3+18AU_1U_2U_3+(3A(U_1)^2-1)U_4+\delta\sqrt{A}(1-(U_1)^2)/2\,. \tag{4.2}$$

The equilibrium points are $U^\pm=\pm(1,0,0,0,0)^T$, and at these points the characteristic polynomials are

$$\mathcal{P}^\pm(\lambda)=\lambda^5+\lambda^3(1-3A)\pm\delta\sqrt{A}\,. \tag{4.3}$$

For small δ the manifolds $W^u(U^+)$ and $W^s(U^-)$ are both two-dimensional, resulting in a codimension two event when searching for heteroclinic solutions connecting the two hyperbolic fixed points U^+ and U^-. The HCCH equation exhibits the same reversibility properties as its lower order version. This reversibility is again given by the transformation (2.8) from the CCH section, which also here fulfills $RF(U)=-F(RU)$. The codimension reduces by one and again we deal with a codimension one problem and two parameters; hence we may expect solution branches in the (A,δ) parameter plane. Section 2.1 showed that a condition for the existence of heteroclinic orbits is a value where the distance function (2.9) reaches zero, and the same condition holds for the HCCH equation. The phase space is sketched in Figure 4.1, indicating the linearizations of the intersecting manifolds in the equilibrium points. For this problem a shooting method will be very slow and may lead to inaccuracy since the additional parameter, say $\varphi\in[0,2\pi)$, an angle defining points on a circle close to the equilibrium point on the linearization of the two-dimensional manifold, requires a very fine resolution to obtain heteroclinic solutions.

4.1. Boundary value problem formulation. There exist several possibilities for setting up equations for finding heteroclinic connections in a boundary value problem framework. Generally one crucial stumbling block is the choice of a suitable phase condition that picks out a certain solution from the infinitely many available ones due to phase shifts [8, 2]. We choose to incorporate one phase condition, proposed by Beyn [2], for which we use an approximation of the solution, V, typically given by a previous solution for slightly different parameter values. Equation (4.1) contains two parameters, A, δ, and in addition the truncated domain length L. As discussed by Doedel and Friedman [5], one of the free parameters can be replaced by L to find a connection. For a nearby chosen and fixed δ, we treat L and A as free parameters. We extrapolate in the (A, δ) plane to get an approximate value of A and solve for the exact data. Rescaling the domain to $[0, 1]$ yields, with the phase condition variable U_{ph} introduced by Beyn [2], the first-order system

$$\text{(4.4a)} \qquad U_i' = LU_{i+1}, \quad i = 1, 2, 3, 4,$$

$$\text{(4.4b)} \qquad U_5' = L\left(6A(U_2)^3 + 18AU_1U_2U_3 + (3A(U_1)^2 - 1)U_4 + \delta\sqrt{A}\frac{(1-(U_1)^2)}{2}\right),$$

$$\text{(4.4c)} \qquad U_{ph}' = L(V')^T U,$$

$$\text{(4.4d)} \qquad L' = 0,\ A' = 0.$$

Hence, we obtain one equation for the phase condition and two for the parameters in addition to the five given by the original ODE; i.e., we have an overall system of eight equations, which have to be supplemented by the same number of boundary conditions. At the edges of the domain we utilize *projected boundary conditions* [4, 2], which make use of eigenvectors in the equilibrium points and can be incorporated by computing V_0, the matrix whose columns are composed of the eigenvectors which correspond to the eigenvalues at the upper equilibrium point U^+ with negative real part, and by forming the counterpart V_1 containing those eigenvectors given by the unstable directions at the lower stationary point U^-. Hence, we consider the eight boundary conditions

$$\text{(4.5)} \qquad U_{ph}(0) = 0, \quad U_{ph}(1) = 0, \quad V_0^T(U(0) - U^+) = 0, \quad V_1^T(U(1) - U^-) = 0\,.$$

For initial estimates we can use solutions obtained from the asymptotic analysis of section 3.1, i.e., the leading order solution tanh profiles

$$V(x) = -\tanh(x - K) + \tanh(x) - \tanh(x + K),$$

for the het_1 solution with guessed root-distance K.

The boundary value solvers we use are based on mono-implicit Runge–Kutta formulae [23, 14]. As for the CCH problem, efficiency can be improved by making use of the theory from section 2.1, which holds analogously for the HCCH equation, to obtain a boundary condition at the fixed point of a point-symmetric solution. We can use half of the previous domain length, and phase conditions become redundant because the phase is already fixed. We replace the projected boundary conditions by

$$U_1(0) = 1, \quad U_2(0)^2 + U_3(0)^2 = 0, \quad U_4(0)^2 + U_5(0)^2 = 0$$

so that together with the self-reversibility condition on the right interval end $U_1(1) = U_3(1) = U_5(1) = 0$ we have six conditions which match the five equations together with the free parameter A. Final solutions are obtained by reflecting the solution and its derivatives around zero and changing the signs of the first, third, and fifth components. Examples of branches of different solutions are shown in Figure 4.2.

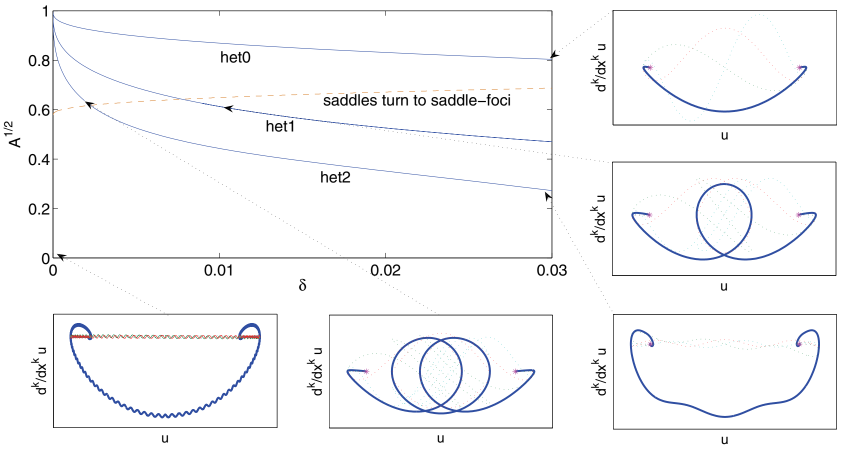


FIG. 4.2. *$(\sqrt{A}, \delta)$-plane with curves for the first three heteroclinic connection branches for the HCCH equation. The dashed line in the parameter plane indicates the position where the positive roots of the characteristic polynomial in U^+ have nonzero imaginary parts. Below and to the right we see five phase-space diagrams (tuples $(U_1, U_2), (U_1, U_3), \dots$) for selected solutions pointed out with arrows marking the corresponding parameters. The first pair (U_1, U_2) is plotted as a bold solid curve.*

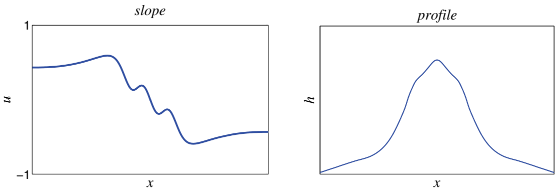


FIG. 4.3. *het_2 solution for $\delta = 0.01$ and $A = 0.443$ and the corresponding profile obtained by integration.*

4.2. Solutions and comparison to analytical results. With the boundary value formulation we are able to compute new HCCH stationary solutions. In Figure 4.3 we see a particular het_2 solution and the profile of the growing structure.

Up to three dimensions one can nicely visualize heteroclinic orbits in the corresponding phase space, while when the dimension is four or higher and the derivatives vanish in the far field one can still plot the two-dimensional phase spaces $(U_1, U_2), (U_1, U_3), \dots$ and demand connections between the equilibrium tuples $(\pm\sqrt{A}, 0)$ as a necessary condition for heteroclinic orbits in the higher order space. Several such projections onto two dimensions are shown in Figure 4.2, where we also see a very rapidly oscillating heteroclinic curve in the bottom left plot, which was found by a shooting approach with a minimization procedure that used the two parameters and an angle as free parameters and the distance function (2.9) as an objective function, depending on those parameters. It indicates that, as shown for the CCH equation, we can in fact find many more het_k branches than those presented for $k = 0, 1, 2$, all emerging from $(A, \delta) = (1, 0)$, which corresponds to the Cahn–Hilliard equation.

In Figure 4.4 we see the change in appearance of solutions on the het_2 branch as δ is increased. The shape varies from a solution with two pronounced humps to a monotone one, similar to the het_0 solution, although associated with different,

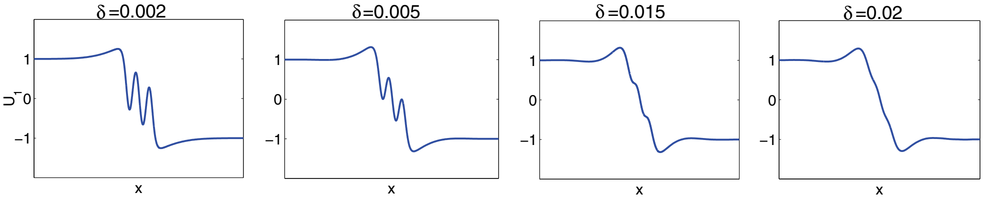


Fig. 4.4. *Structural change of the scaled het$_2$ solution as δ is increased.*

smaller values of A. This is crucial if one wants to compute solutions for bigger δ with a boundary value solver. It easily happens that the solver switches between solution branches; however, this can be prevented by starting continuation in a parameter regime where the high-slope parts of the solutions are nonmonotone, and continuing with small steps. A characteristic of the HCCH solutions is the overshoot from the equilibrium value before the solutions tend into the direction of the negative equilibrium point. This is not observed for the CCH equation, where the shape is similar at these regions to hyperbolic tangent functions.

In light of the expansion (3.2) we try to estimate the $\mathcal{O}(\delta^{1/3})$ terms A_1 for the different heteroclinic connections in a range of very small δ. As we see in Figure 4.5 on the left, the numerically obtained values for A behave like $A = 1 - 2^{1/6}\delta^{1/3}$ in the case of the het_0 solutions, so that $A_1 = -2^{1/6}$, which is consistent with the result in Savina et al. [22]. The numerical result for het_1 is in line with the analytical value (3.62), and since for het_2 we see the agreement $A_1 \approx -5\,2^{1/6}$, we propose for higher order trajectories that for het_k we have the general approximation $A_1 \approx -(2k+1)\,2^{1/6}$, which is reminiscent of the CCH expression (2.45). Hence this formula is used in Figure 4.5 to plot the *analytical values.*

We measure the distance between the first and second roots for the het_1 and het_2 solutions, as seen in Figure 4.5 on the right. We compare this to the analytical expression (3.61) for the 1-hump solutions in the same figure and see that for small δ the agreement is good. For both het_1 and het_2 solutions the distance is seen to increase logarithmically as δ decreases.

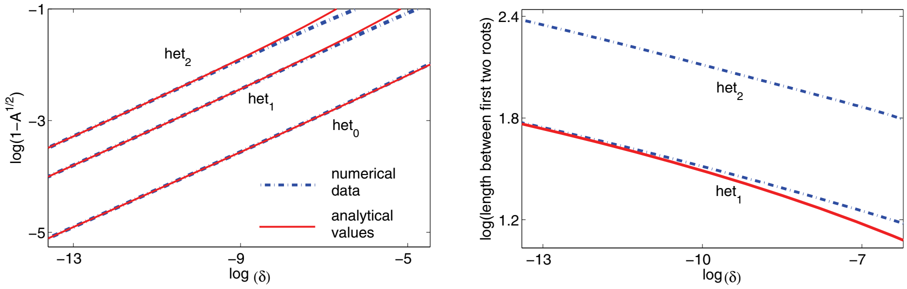


Fig. 4.5. *Left figure: Logarithmic version of the* $(\sqrt{A}, \delta)$ *plot for very small* δ*. The solid lines give the analytical values, and the dash-dotted lines give those computed with the BVP solver. On the right we see the distances between the first two roots of the het*$_1$ *and het*$_2$ *solutions numerically and for het*$_1$ *via the analytical expression* (3.61) *(solid line).*

5. Conclusion.

We have demonstrated that a sixth-order generalization of the CCH equation admits multiple stationary solutions connecting constant values. As for the fourth-order CCH solution, these include a simple base solution, which is monotone for the CCH and "almost" monotone for the HCCH equation. More complex solutions, containing multiple humps, are also possible for each value of the forcing parameter δ, given particular values of the integration constant $A(\delta)$. These nonmonotone stationary solutions constitute an essential part of the solution structure for this higher order Cahn–Hilliard type of equation. We have demonstrated this via a numerical investigation of the phase space in which we are able to follow solution branches. For the simplest of the multi-humped solutions, the het_1 branch, careful use of matched asymptotics that accounts for exponentially small terms allows us to find a solution which yields both the length scale for the solution (the "hump length") and the parameter value $A(\delta)$, at which it occurs, in the limit of small δ. Extension of the analysis to higher branches appears feasible. Our numerical evidence suggests that similarly simple asymptotic expressions hold for these branches for both the CCH and HCCH equations.

Various issues, such as the stability of these solutions, are presently being considered in light of applications of the HCCH equation as a model for the morphology and dynamics of quantum dots. In particular, how do adjacent internal layers derived from these solutions interact, and what is their effect on the coarsening behavior in large spatial domains? Savina et al. [22] have begun an investigation of these questions by numerical simulation of (1.4); it is likely that asymptotics can yield further insights.

Physically, further interesting questions relate to the extension of the HCCH model to richer models for the energetics of faceted surfaces, and analyzing the three-dimensional extension of the model.

Appendix A. Outer problem. For the solution to the outer problem (3.6), (3.7) it is easy to observe that to leading order in δ the solution of

$$Q\left(Y_0\right)_{\xi\xi\xi} = -\sqrt{2}\left(Y_0^2 - 1\right) \quad \text{with} \quad \lim_{\xi\to-\infty} Y_0(\xi) = 1 \tag{A.1}$$

is

$$Y_0(\xi) = 1\,. \tag{A.2}$$

To $O(\delta^{1/3})$ the general solution to the problem

$$Y_{1_{\xi\xi\xi}} - \sqrt{2}\,Y_1 = -\frac{A_1}{\sqrt{2}} \quad \text{with} \quad \lim_{\xi\to-\infty} Y_1(\xi) = \frac{A_1}{2} \tag{A.3}$$

is

$$Y_1(\xi) = \frac{A_1}{2} + C_1 e^{2^{1/6}\xi} + e^{-\xi/2^{5/6}}\left[C_2\cos\left(\sqrt{3}\,\xi/2^{5/6}\right) + C_3\sin\left(\sqrt{3}\,\xi/2^{5/6}\right)\right], \tag{A.4}$$

with C_1, C_2, and C_3 being constants of integration. The far-field condition requires that Y_1 remains bounded as $\xi \to -\infty$. Hence, $C_2 = C_3 = 0$ and

$$Y_1(\xi) = \frac{A_1}{2} + C_1 e^{2^{1/6}\xi}. \tag{A.5}$$

Using this and the far-field conditions, the solution to the $O(\delta^{2/3})$ problem

$$Y_{2_{\xi\xi\xi}} - \sqrt{2}\,Y_2 = -\frac{A_2}{\sqrt{2}} - \frac{1}{2}\left(3\left(Y_1^2\right)_{\xi\xi\xi} - \sqrt{2}\,Y_1^2\right) \quad \text{with} \quad \lim_{\xi\to-\infty} Y_2(\xi) = \frac{A_2}{2} - \frac{A_1^2}{8} \tag{A.6}$$

is

$$Y_2(\xi) = \frac{A_2}{2} - \frac{A_1^2}{8} + D_1 e^{2^{1/6}\xi} + \frac{A_1 C_1}{3} e^{2^{1/6}\xi}\left(1 - 2^{1/6}\xi\right) - \frac{23}{14} C_1^2 e^{2^{7/6}\xi}, \tag{A.7}$$

and the solution to the $O(\delta)$ problem

$$Y_{3_{\xi\xi\xi}} - \sqrt{2}\,Y_3 = -\frac{A_2}{\sqrt{2}} + \frac{Y_{1_{\xi\xi\xi\xi}}}{4} + \sqrt{2}\,Y_1 Y_2 - \frac{1}{2}\left(Y_1^3 + 6\,Y_1 Y_2\right)_{\xi\xi\xi}$$
$$\text{with} \quad \lim_{\xi\to-\infty} Y_3(\xi) = \frac{A_3}{2} - \frac{A_1 A_2}{4} + \frac{A_1^3}{16} \tag{A.8}$$

is

(A.9)
$$\begin{aligned} Y_3(\xi) = {} & \frac{A_3}{2} - \frac{A_1 A_2}{4} + \frac{A_1^3}{16} \\ & + \Bigg[K_1 - \frac{2^{1/3}}{12} C_1 + \frac{1}{3}\left(A_1 D_1 + A_2 C_1\right) - \frac{59}{216} C_1 A_1^2 \\ & + \left(\frac{\sqrt{2}}{12} C_1 - \frac{2^{1/6}}{3}\left(A_1 D_1 + A_2 C_1\right) + \frac{17}{72} 2^{1/6} A_1^2 C_1 \right)\xi + \frac{2^{1/3}}{18} A_1^2 C_1\,\xi^2 \Bigg] e^{2^{1/6}\xi} \\ & + \left[-\frac{23}{7} C_1 D_1 + \left(\frac{7}{12} + \frac{23}{21} 2^{1/6}\xi\right) A_1 C_1^2 \right] e^{2^{7/6}\xi} + \frac{127}{28} C_1^3 e^{2^{1/6} 3\xi}\,, \end{aligned}$$

with another integration constant K_1. Finally, we obtain the asymptotic representation in terms of x_m:

(A.10)
$$\begin{aligned} Y(x_m) = 1 & + \left[C_1 + \frac{1}{2} A_1 \right] \delta^{1/3} \\ & + \left[C_1\, 2^{1/6}\, x_m - \frac{1}{8}\, A_1^2 + \frac{1}{3}\, C_1\, A_1 + D_1 - \frac{23}{14} C_1^2 + \frac{1}{2}\, A_2 \right] \delta^{2/3} \\ & + \Bigg[-\frac{23}{7} C_1^2\, 2^{1/6} x_m + D_1\, 2^{1/6} x_m + \frac{1}{2}\, C_1\, 2^{1/3} x_m^2 + \left(-\frac{1}{4}\, A_1 + \frac{1}{3}\, C_1 \right) A_2 \\ & \quad + \left(\frac{7}{12} C_1^2 + \frac{1}{3}\, D_1 \right) A_1 + \frac{1}{2}\, A_3 - \frac{59}{216} C_1\, A_1^2 - \frac{1}{12}\, 2^{1/3}\, C_1 \\ & \quad + K_1 - \frac{23}{7} C_1\, D_1 + \frac{1}{16} A_1^3 + \frac{127}{28} C_1^3 \Bigg]\ \delta. \end{aligned}$$

REFERENCES

[1] K. L. Adams, J. R. King, and R. H. Tew, *Beyond-all-orders effects in multiple-scales asymptotics: Travelling-wave solutions to the Kuramoto–Sivashinsky equation*, J. Engr. Math., 45 (2003), pp. 197–226.

[2] W.-J. Beyn, *The numerical computation of connecting orbits in dynamical systems*, IMA J. Numer. Anal., 9 (1990), pp. 379–405.

[3] G. Carrier and C. Pearson, *Ordinary Differential Equations*, Blaisdell, Waltham, MA, 1968.

[4] F. R. de Hoog and R. Weiss, *An approximation theory for boundary value problems on infinite intervals*, Computing, 24 (1980), pp. 227–239.

[5] E. J. Doedel and M. J. Friedman, *Numerical computation of heteroclinic orbits*, J. Comput. Appl. Math., 26 (1989), pp. 155–170.

[6] A. Eden and V. K. Kalantarov, *The convective Cahn–Hilliard equation*, Appl. Math. Lett., 20 (2007), pp. 455–461.

[7] C. L. Emmott and A. J. Bray, *Coarsening dynamics of a one-dimensional driven Cahn–Hilliard system*, Phys. Rev. E, 54 (1996), pp. 4568–4575.

[8] M. J. Friedman and E. J. Doedel, *Numerical computation and continuation of invariant manifolds connecting fixed points*, SIAM J. Numer. Anal., 28 (1991), pp. 789–808.

[9] A. A. Golovin, S. H. Davis, and A. A. Nepomnyashchy, *A convective Cahn-Hilliard model for the formation of facets and corners in crystal growth*, Phys. D, 122 (1998), pp. 202–230.

[10] A. A. Golovin, A. A. Nepomnyashchy, S. H. Davis, and M. A. Zaks, *Convective Cahn-Hilliard models: From coarsening to roughening*, Phys. Rev. Lett., 86 (2001), pp. 1550–1553.

[11] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Clarendon Press, Oxford, UK, 1993.

[12] C. J. Howls, T. Kawai, and Y. Takei, eds., *Toward the Exact WKB Analysis of Differential Equations, Linear or Nonlinear*, Kyoto University Press, Kyoto, 1999.

[13] W. Kath, C. Knessl, and B. Matkowsky, *A variational approach to nonlinear singularly perturbed boundary value problems*, Stud. Appl. Math., 77 (1987), pp. 61–88.

[14] J. Kierzenka and L. Shampine, *A BVP solver based on residual control and the MATLAB PSE*, ACM Trans. Math. Software, 27 (2001), pp. 299–316.

[15] C. G. Lange, *On spurious solutions of singular perturbation problems*, Stud. Appl. Math., 68 (1983), pp. 227–257.

[16] K.-T. Leung, *Theory on morphological instability in driven systems*, J. Statist. Phys., 61 (1990), pp. 345–364.

[17] F. Liu and H. Metiu, *Dynamics of phase separation of crystal surfaces*, Phys. Rev. B, 48 (1993), pp. 5808–5817.

[18] R. E. O'Malley, Jr., *Phase-plane solutions to some singular perturbation problems*, J. Math. Anal. Appl., 54 (1976), pp. 449–466.

[19] L. G. Reyna and M. J. Ward, *Metastable internal layer dynamics for the viscous Cahn-Hilliard equation*, Methods Appl. Anal., 2 (1995), pp. 285–306.

[20] S. Rosenblat and R. Szeto, *Multiple solutions of nonlinear boundary-value problems*, Stud. Appl. Math., 63 (1980), pp. 99–117.

[21] Y. Saito and M. Uwaha, *Anisotropy effect on step morphology described by Kuramoto-Sivashinsky equation*, J. Phys. Soc. Japan, 65 (1996), pp. 3576–3581.

[22] T. V. Savina, A. A. Golovin, S. H. Davis, A. A. Nepomnyashchy, and P. W. Voorhees, *Faceting of a growing crystal surface by surface diffusion*, Phys. Rev. E, 67 (2003), 021606.

[23] L. F. Shampine, P. H. Muir, and H. Xu, *A user-friendly Fortran BVP-solver*, JNAIAM J. Numer. Anal. Ind. Appl. Math., 1 (2006), pp. 201–217.

[24] V. A. Shchukin and D. Bimberg, *Spontaneous ordering of nanostructures on crystal surfaces*, Rev. Modern Phys., 71 (1999), pp. 1125–1171.

[25] M. J. Ward, *Eliminating indeterminacy in singularly perturbed boundary value problems with transition invariant potentials*, Stud. Appl. Math., 87 (1992), pp. 95–134.

[26] S. J. Watson, F. Otto, B. Rubinstein, and S. H. Davis, *Coarsening dynamics of the convective Cahn-Hilliard equation*, Phys. D, 178 (2003), pp. 127–148.

[27] C. Yeung, T. Rogers, A. Hernandes-Machado, and D. Jasnow, *Phase separation dynamics in driven diffusive systems*, J. Statist. Phys., 66 (1992), pp. 1071–1088.

[28] M. A. Zaks, A. Podolny, A. A. Nepomnyashchy, and A. A. Golovin, *Periodic stationary patterns governed by a convective Cahn–Hilliard equation*, SIAM J. Appl. Math., 66 (2006), pp. 700–720.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 375–397

MESENCHYMAL MOTION MODELS IN ONE DIMENSION[*]

ZHI-AN WANG[†], THOMAS HILLEN[‡], AND MICHAEL LI[‡]

Abstract. Mesenchymal motion denotes a form of cell movement through tissue which can be observed for certain cancer metastases. In [T. Hillen, *J. Math. Biol.*, 53 (2006), pp. 585–616], a mathematical model for this form of movement was introduced. In the current paper we present a comprehensive analysis of the one-dimensional mesenchymal motion model. We establish the global existence of classical solutions and rigorously carry out the parabolic limit of the model. We discuss the stationary solutions, prove the existence of traveling wave solutions, and use numerical simulations to illustrate the results. Finally, we discuss the biological implications of the results.

Key words. mesenchymal motion, stationary solutions, global existence, macroscopic limits, traveling waves, hyperbolic systems

AMS subject classifications. 35L45, 35L60, 92B99

DOI. 10.1137/080714178

1. Introduction.

Mesenchymal motion is a form of cellular movement through tissue which is formed from fiber networks. An example is the invasion of tumor metastases through collagen networks [7]. Cells migrate in fiber networks and change their directions according to the orientational distribution of fibers. Moreover, cells actively remodel the matrix by excreting a matrix degrading enzyme (e.g., protease) to generate sufficient space in which to migrate.

The motion of mesenchymal cells in a tissue matrix was reported in a review article by Friedl and Bröcker [7]. Mesoscopic and macroscopic mathematical models for mesenchymal motion were derived by Hillen [12] in a temporally varying network tissue. The mesoscopic models consist of a transport equation for the cell movement and an ordinary differential equation (ODE) for the dynamics of tissue fibers. The macroscopic models have the form of drift-diffusion equations, where the mean drift velocity is given by the mean orientation of the tissue, and the diffusion tensor is given by the variance-covariance matrix of the tissue orientation. The analysis in [12] is divided into the case of undirected and directed tissue according to the distribution of fiber orientation. In undirected tissue, the fibers are symmetrical along their axes and both fiber directions are identical. For example, collagen fibers are undirected and form the basis for many human and animal tissues. For directed tissue, the fibers are unsymmetrical and the two ends can be distinguished (such as microtubules and actin filaments). Branching collagen fiber networks can also be considered directional if the branching points are of significance for the movement of cells [12].

The model from [12] was extended in [3, 4] to include cell-cell interactions and chemotactic forces for the case of undirected fibers. Formal methods were used to derive the corresponding macroscopic models. Painter [21] numerically studied models for cell movement in fiber tissues and showed pattern formation in the form of

[*] Received by the editors January 24, 2008; accepted for publication (in revised form) June 11, 2008; published electronically November 5, 2008.

http://www.siam.org/journals/siap/69-2/71417.html

[†] Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN 55455 (zhiwang@ima.umn.edu). This author's work was supported by the IMA postdoctoral fellowship.

[‡] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada (thillen@math.ualberta.ca, mli@math.ualberta.ca). The work of these authors was supported by the Natural Science and Engineering Council of Canada (NSERC).

macroscopic networks.

In this paper, the one-dimensional mesenchymal motion model is fully analyzed. The global existence of solutions, macroscopic limits, traveling waves, and stationary solutions are investigated. The one-dimensional model is very instructive and we can gain much insight into the mechanisms involved in the model. For example, we find the existence of traveling pulse solutions for the cell population and identify some mechanisms for cell aggregation. We also identify some differences between undirected and directed tissue by analyzing the one-dimensional model. We restrict our attention to the model for directed tissue only and the analysis can be completely adopted to the study for undirected tissue from the mathematical point of view.

The paper is organized as follows. In the rest of this section, we will present the one-dimensional mesenchymal motion model derived in [12] and discuss the stationary solutions based on the telegraph process analysis. In section 2, we classify the one-dimensional model as a degenerated hyperbolic system and conclude that there is no shock solution. In section 3, the global existence of classical solutions is obtained along the characteristics using a fixed point argument and general regularity results for the semilinear hyperbolic system. In section 4, we rigorously carry out the parabolic limit of the one-dimensional mesenchymal transport model, where we show that solutions of the one-dimensional model converge to solutions of the corresponding drift-diffusion limit equation. In section 5, we study the traveling wave solutions and find traveling pulse solutions for the cell population and traveling front waves for fiber orientations. In the final section 6, we summarize and compare our results with the results obtained in [12]. Furthermore, we explain the findings in the context of the biological application of cell movement in tissue.

1.1. Models for mesenchymal motion in one dimension. In this paper, we are primarily interested in the one-dimensional mesenchymal motion model for the case of directed tissue, which reads as follows [12]:

$$
\begin{aligned}
p_t^+ + sp_x^+ &= -\mu p^+ + \mu q^+(p^+ + p^-),\\
p_t^- - sp_x^- &= -\mu p^- + \mu q^-(p^+ + p^-),\\
q_t^+ &= \kappa(p^+ - p^-)(q^- - q^+ + 1)q^+,\\
q_t^- &= \kappa(p^+ - p^-)(q^- - q^+ - 1)q^-.
\end{aligned}
\tag{1.1}
$$

The quantities p^+, p^- denote density of cells moving to the right or left, respectively, with a constant speed s. The functions q^+, q^- are distributions of fibers pointing to the right $(+)$ or left $(-)$. The constant $\mu \geq 0$ denotes the turning rate, and the constant $\kappa \geq 0$ represents the cutting efficiency (rate of fiber degradation). The transport term $sp_x^\pm$ in (1.1) accounts for the cell migration in either direction with speed s. The right-hand side of the first two equations describes the change of cell movement in the field of fibers. The third and fourth equations of (1.1) describe the changes of the fibers in either direction due to the interaction with cells. The derivation of the above model is omitted here for brevity and we refer interested readers to [12] for details. It is worthwhile to point out that the model for undirected tissue can be regarded as a special case of (1.1) for $\kappa = 0$ (see also [12]). In this paper, we focus on the model of directed tissue, and most of our results can be applied to the case of undirected tissue. The significant difference, when it appears, will be emphasized.

The system (1.1) is closely related to the Goldstein–Kac system [8, 17] which describes correlated random walk in one space dimension. With $p = p^+ + p^-$,

$j = s(p^+ - p^-)$, $q = q^+ + q^-$, and $\xi = q^+ - q^-$, system (1.1) becomes

$$\begin{aligned} p_t + j_x &= \mu(q-1)p, \\ j_t + s^2 p_x &= -\mu j + \mu s \xi p, \\ q_t &= (\kappa/s) j \xi (1-q), \\ \xi_t &= (\kappa/s) j (q - \xi^2). \end{aligned} \tag{1.2}$$

Since (q^+, q^-) denotes a distribution, $q^+ + q^- = 1$. Hence one is interested in solutions with $q = 1$. The set $q = 1$ is an invariant manifold of the system (1.2) which will be verified later in Lemma 3.1. On this manifold the system (1.2) reduces to

$$\begin{aligned} p_t + j_x &= 0, \\ j_t + s^2 p_x &= -\mu j + \mu s \xi p, \\ \xi_t &= (\kappa/s) j (1 - \xi^2). \end{aligned} \tag{1.3}$$

If q^+ is used instead of ξ, then the system becomes

$$\begin{aligned} p_t + j_x &= 0, \\ j_t + s^2 p_x &= -\mu j + \mu s (2q^+ - 1)p, \\ q_t^+ &= 2(\kappa/s) j q^+ (1 - q^+). \end{aligned} \tag{1.4}$$

Finally, if the Kac's trick is applied to the first two equations of the above equation, then a damped wave equation is obtained:

$$p_{tt} + \mu p_t = s^2 p_{xx} - \mu s((2q^+ - 1)p)_x. \tag{1.5}$$

Any of (1.1)–(1.5) will be used for a particular question as shown later.

Now we investigate the connections between the one-dimensional mesenchymal motion model and the well-known Goldstein–Kac model [8, 17]. We use the normalization condition $q^+ + q^- = 1$ to substitute $q^- = 1 - q^+$ into the first two equations of (1.1) and obtain

$$\begin{aligned} p_t^+ + s p_x^+ &= -\mu(1 - q^+)p^+ + \mu q^+ p^-, \\ p_t^- - s p_x^- &= \mu(1 - q^+)p^+ - \mu q^+ p^-. \end{aligned} \tag{1.6}$$

The model for the case of undirected tissue ($\kappa = 0$) possesses some very interesting properties. Undirected tissue fibers are symmetrical along their axes and both fiber directions are identical, which indicates that $q^+ = q^- = \frac{1}{2}$. Then the model (1.6) becomes the Goldstein–Kac model [8, 17]

$$\begin{aligned} p_t^+ + s p_x^+ &= \frac{\mu}{2}(p^- - p^+), \\ p_t^- - s p_x^- &= -\frac{\mu}{2}(p^- - p^+). \end{aligned} \tag{1.7}$$

The parabolic scaling for the Goldstein–Kac model, which leads to a parabolic equation, has been discussed in [9] and references therein.

For directed tissue, we define $\lambda^+ = \mu(1 - q^+), \lambda^- = \mu q^+$; then (1.6) is converted into

$$\begin{aligned} p_t^+ + s p_x^+ &= -\lambda^+ p^+ + \lambda^- p^-, \\ p_t^- - s p_x^- &= \lambda^+ p^+ - \lambda^- p^-, \end{aligned} \tag{1.8}$$

which is a modification of the Goldstein–Kac model. Extensions of the Goldstein–Kac model and local and global existence of the solution to the extended model have been extensively investigated in the literature [14, 15, 16]. The telegraph process of (1.8) has been briefly discussed recently by Erban and Othmer [5]. The results obtained in [14, 15, 16] can be applied to system (1.8) if the turning rates $\lambda^{\pm}(t,x)$ are given functions. The theory does not, however, apply to (1.1), since the turning rates are coupled with the $q^{\pm}$ equations.

In the next subsection, we will discuss stationary solutions for (1.1) based on the telegraph process examined in [12].

We supply the system (1.1) with the initial condition

$$p^{\pm}(0,x) = p_I^{\pm}(x), \quad q^{\pm}(0,x) = q_I^{\pm}(x), \quad x \in \Omega. \tag{1.9}$$

Due to the biological interest and normalization condition $q^+ + q^- = 1$, we make the following assumptions for the initial data and boundary conditions.

(ic) $p_I^{\pm} \geq 0$, $0 \leq q_I^+, q_I^- \leq 1$, and $q_I^+ + q_I^- = 1$. For undirected tissue, we assume that the initial data is symmetrical, i.e., $q_I^+ = q_I^- = \frac{1}{2}$.

Here we consider two types of boundary conditions.

(bc1) $\Omega = \mathbb{R}$ and $p_I^{\pm}(x), q_I^{\pm}(x)$ have compact support in Ω.

(bc2) $\Omega = [-l, l]$ and zero flux boundary condition, namely,

$$p^+(t, \pm l) = p^-(t, \pm l).$$

1.2. Stationary solutions. In this section we discuss stationary solutions of the mesenchymal transport model (1.1) using an argument similar to that in [6]. We first present a second-order telegraph equation which is derived from system (1.1). To this end, we add and subtract the first two equations of (1.1) and obtain equations for the total population $p = p^+ + p^-$ and the population flux $j = s(p^+ - p^-)$,

$$\begin{aligned} p_t + j_x &= 0, \\ j_t + s^2 p_x &= -\mu j + \mu(q^+ - q^-)sp, \end{aligned} \tag{1.10}$$

with initial conditions $p(0,x) = p_I(x)$ and $j(0,x) = j_I(x)$, where p_I and j_I are determined from the initial condition (1.9) of p^+ and p^-. We differentiate the first equation of (1.10) with respect to t and the second equation with respect to x. After that, we subtract the resulting equations and end up with a damped wave equation with drift term (see (1.5) or [12])

$$p_{tt} + \mu p_t + \mu(s\xi_q p)_x = s^2 p_{xx}, \tag{1.11}$$

where the drift velocity is given by the expectation of q denoted by $\xi_q = q^+ - q^-$. Equation (1.11) is a form of telegraph equation which describes electrical transmission in a telegraph cable when current leaks to the ground. A drift-diffusion equation can be approximated by taking the limit $\mu \to \infty, s \to \infty$ with diffusivity $D = s^2/\mu < \infty$ and drift velocity $s\xi_q < \infty$. The same drift-diffusion equation also can be obtained by multiscale methods (see [12]).

Suppose that equations (1.10) are defined in the interval $\Omega = [-l, l]$ and satisfy the boundary condition (bc2). In terms of cell population density, the zero flux boundary condition is equivalent to $p^+(\pm l) = p^-(\pm l) = \frac{1}{2}p(\pm l)$. We want to know under what conditions, if any, these equations have time-independent, space-dependent solutions for $p^{\pm}$. The steady state condition $j_x = 0$ of the first equation of (1.10) implies that

j is a constant, and the zero flux boundary condition $j(\pm l) = 0$ furthermore gives that $j = 0$. Consequently the second equation of (1.10) becomes

$$p_x = \frac{\mu}{s}(q^+ - q^-)p.$$

This is a first-order equation for p, whose solution can be easily found:

$$p(x) = p(-l)\exp\left(\frac{\mu}{s}\int_{-l}^{x}(q^+(y) - q^-(y))dy\right). \tag{1.12}$$

The vanishing flux $j = 0$ gives that $p^+ = p^-$, and hence

$$p^\pm(x) = \frac{p(-l)}{2}\exp\left(\frac{\mu}{s}\int_{-l}^{x}(q^+(y) - q^-(y))dy\right). \tag{1.13}$$

Note that the above integrals are bounded since q^+ and q^- are bounded by 1, which will be proved in section 3. From the above equations, one can see how the distribution of fiber orientations $q^\pm$ affects the distribution of cell populations p and $p^\pm$. In particular, if $\mu \neq 0$ and $q^+ \neq q^-$, then p and $p^\pm$ are nonconstants which correspond to the stationary solutions of the system (1.10).

Particularly in undirected tissue, $q^+ = q^- = \frac{1}{2}$ due to symmetry; then p and $p^\pm$ are constants and $p^+ = p^- = \frac{p(-l)}{2}$, which means that there is no aggregation of cells.

If $q^+ = 1, q^- = 0$, then

$$p^\pm(x) = \frac{p(-l)}{2}\exp\left(\frac{\mu}{s}(x + l)\right).$$

The cells accumulate at the end $x = l$. This is not surprising because all cells bias their movement to the right and eventually accumulate at the right end due to the zero flux boundary condition.

Similarly, if $q^+ = 0, q^- = 1$, then

$$p^\pm(x) = \frac{p(-l)}{2}\exp\left(-\frac{\mu}{s}(x + l)\right),$$

and $p^\pm$ attains the maximum at $x = -l$.

Therefore, here we identify a mechanism which can lead to aggregation, namely, $\mu \neq 0$, and the tissues are directed and cells have a probability 1 moving to the left or right.

2. Classification as hyperbolic system.

In this section we show that the system (1.1) is degenerately hyperbolic, and we discuss shock solutions. To this end, we rewrite (1.1) in a matrix form

$$u_t + \Theta u_x = H(u), \tag{2.1}$$

where u, Θ, and $H(u)$ are defined as

$$u = \begin{bmatrix} p^+ \\ p^- \\ q^+ \\ q^- \end{bmatrix}, \quad \Theta = \begin{bmatrix} s & 0 & 0 & 0 \\ 0 & -s & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad H(u) = \begin{bmatrix} -\mu p^+ + \mu q^+(p^+ + p^-) \\ -\mu p^- + \mu q^-(p^+ + p^-) \\ \kappa(p^+ - p^-)(q^- - q^+ + 1)q^+ \\ \kappa(p^+ - p^-)(q^- - q^+ - 1)q^- \end{bmatrix}.$$

The drift term is linear, and hence the system (2.1) cannot create shock solutions. The 4×4 matrix Θ has eigenvalues $\lambda_1 = -s < 0, \lambda_2 = \lambda_3 = 0, \lambda_4 = s$ satisfying $\lambda_1 < \lambda_2 = \lambda_3 < \lambda_4$ provided that $s > 0$. This implies that the system (2.1) and hence (1.1) are hyperbolic but not strictly hyperbolic since the two eigenvalues λ_2 and λ_3 are identical. The eigenvectors r_i corresponding to eigenvalues λ_i, $i = 1, 2, 3, 4$, are

$$r_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ r_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \ r_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ r_4 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

It can be verified that $\nabla\lambda_i(u) \cdot r_i(u) = 0$ for $i = 1, 2, 3, 4$, where $\nabla\lambda_i(u) \cdot r_i(u)$ means the directional derivative of the eigenvalues λ_i in the direction of the eigenfunction r_i. Hence all characteristic fields (λ_i, r_i) are linearly degenerate [2, 19]. Thus a shock which separates intersecting characteristics defining a discontinuity does not exist. However, the solution might contain contact discontinuities if data are discontinuous (see [2]).

The characteristic slopes are determined from the eigenvalues of the 4×4 matrix Θ in (2.1) by $\frac{dx}{dt} = \lambda_i$, which is never infinite, so the line $t = 0$ is nowhere tangent to a characteristic. Therefore, if initial data for p^+, p^-, q^+, and q^- are given along the line $t = 0$, the resulting Cauchy problem should be well-posed, as shown in the subsequent section.

3. Global existence.

In this section, we will prove the global existence of solutions to the system (1.1) subject to the initial condition (ic) and boundary condition (bc1). For a bounded domain, the analysis for global existence will be a little bit more complicated than for an unbounded domain, due to the boundary conditions, and is left open here.

The system (1.1) is a coupling of two partial differential equations (PDEs) and two ODEs. To prove the global existence of solutions to the system (1.1), we first prove the nonnegativity of solutions.

Lemma 3.1. *Let $p_I^\pm \geq 0$ and $q_I^\pm \geq 0$ with $q_I^+ + q_I^- = 1$. Assume that $p^\pm, q^\pm \in L^\infty(0, T; L^\infty(\mathbb{R}))$ is a solution to system* (1.1) *for some $T > 0$; then $p^\pm \geq 0$ and $0 \leq q^\pm(t, x) \leq 1$ with $q^+ + q^- = 1$.*

Proof. We first show that $q^+ + q^- = 1$. Toward this end, we consider $q = q^+ + q^-$ and $\xi = q^+ - q^-$. Then we add and subtract the third and fourth equations of (1.1) to obtain equations for q and ξ as follows:

$$\begin{aligned} q_t &= -\kappa(p^+ - p^-)(q - 1)\xi, \\ \xi_t &= \kappa(p^+ - p^-)(q - \xi^2), \end{aligned} \tag{3.1}$$

which can be rewritten in vector form

$$Q_t = -\kappa(p^+ - p^-)F(Q), \tag{3.2}$$

where

$$Q = \begin{pmatrix} q \\ \xi \end{pmatrix}, \qquad F(Q) = \begin{pmatrix} (q-1)\xi \\ \xi^2 - q \end{pmatrix}.$$

The initial data of the system (3.1) is given by

$$q_I = q_I^+ + q_I^- = 1, \qquad \xi_I = q_I^+ - q_I^-. \tag{3.3}$$

It is straightforward to verify that the vector field $F(Q) \in C^1(\mathbb{R}^2)$, and hence it is locally Lipschitz continuous with respect to Q for a given $p^{\pm} \in L^\infty(0,T;L^\infty(\mathbb{R}))$. Then the Cauchy problem (3.1), (3.3) has a unique solution by the fundamental existence-uniqueness theorem. On the other hand, it is trivial to check that $q = 1$ is a solution of the first equation of (3.1) satisfying initial condition (3.3). Hence the system (3.1), (3.3) has a unique solution $(q = 1, \xi)$, where ξ is determined by the equation

$$\xi_t = \kappa(q^+ - q^-)(1 - \xi^2), \qquad \xi_I = q_I^+ - q_I^-.$$

It is worthwhile to point out that we provide an idea here for proving that $q = 1$ and for proving the (local) existence of q and ξ given that $p^{\pm} \in L^\infty(0,T;L^\infty(\mathbb{R}))$. This idea will be used later without repeating this procedure.

We proceed to show that solutions $q^{\pm}$ preserve the positivity. Substituting $q^- = 1 - q^+$ into the third equation of (1.1), we have

$$q_t^+ = 2\kappa(p^+ - p^-)(1 - q^+)q^+. \tag{3.4}$$

There are three cases to consider.

Case 1. $q_I^+ = 1$. Then we conclude that $q^+ = 1$ is a solution to (3.4) with initial condition $q_I^+ = 1$. Since the right-hand side of (3.4) is locally Lipschitz continuous with respect to q^+, the solution of (3.4) is unique. Hence $q^+(t,x) = 1$ for all t, x.

Case 2. $q_I^+ = 0$. Using an argument similar to Case 1 we can show that $q^+(t,x) = 0$ is a unique solution to (3.4).

Case 3. $0 < q_I^+ < 1$. Then integrating (3.4) with respect to t from 0 to t, one has

$$\frac{q^+}{1 - q^+} = \frac{q_I^+}{1 - q_I^+} \exp\left(\int_0^t 2\kappa(p^+(\tau,\cdot) - p^-(\tau,\cdot))d\tau\right).$$

Due to $0 < q_I^+ < 1$, we have

$$\frac{q^+}{1 - q^+} \geq 0.$$

It follows immediately from the above equality that $0 \leq q^+ \leq 1$. Combining Cases 1, 2, and 3, we get that $0 \leq q^+ \leq 1$ for $0 \leq q_I^+ \leq 1$. Applying $q^+ = 1 - q^-$ in the fourth equation of (1.1) and using the same approach, we can show that $0 \leq q^- \leq 1$.

Finally, we show the positivity of cell density $p^{\pm}(t,x)$. We use the theory of invariant principle from [11] for the hyperbolic random walk system to achieve this goal. To this end, we write the first two equations of the system (1.1) in a matrix form

$$\phi_t = G\phi + B\phi + \mathcal{F}(\phi), \tag{3.5}$$

where

$$\phi = \begin{pmatrix} p^+ \\ p^- \end{pmatrix}, \ G = \begin{pmatrix} -s\dfrac{\partial}{\partial x} & 0 \\ 0 & s\dfrac{\partial}{\partial x} \end{pmatrix}, \ B = \begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix},$$

and

$$\mathcal{F}(\phi) = \begin{pmatrix} \mu q^+(p^+ + p^-) - \mu p^- \\ \mu q^-(p^+ + p^-) - \mu p^+ \end{pmatrix}.$$

Let $\Lambda = [0, \infty) \subset \mathbb{R}$. Then Λ is convex, and for each $z \in \partial\Lambda$, Λ has an outward normal vector. Moreover, define $\Sigma = \Lambda \times \Lambda$. Let $\phi \in \partial\Sigma$, and without loss of generality we assume that $\phi = (\vartheta, 0)$ with $\vartheta \geq 0$. Then for the outward normal vector $\eta(\phi) = (0, -1)$ of ϕ, we have

$$\eta(\phi) \cdot (B\phi + \mathcal{F}(\phi)) = -\mu q^- \vartheta \leq 0,$$

where we have used the positivity of q^-. Then by the theory in [11, Theorem 2], the set Σ is positively invariant for the system (3.5), which shows the positivity of $p^\pm$. The proof is completed. □

By Lemma 3.1, we obtain the following theorem.

Theorem 3.2. *The set* $\{\ (p^+, p^-, q^+, q^-) \mid p^\pm \geq 0, q^\pm \geq 0, q^+ + q^- = 1\}$ *is invariant to the system* (1.1) *provided that* $p^\pm, q^\pm \in L^\infty(0, T; L^\infty(\mathbb{R}))$ *for* $T > 0$.

Remark 1. For $p^+ > p^-$, the term $p^+ - p^- > 0$ and q^+ will increase while q^- decreases. Hence directionality is enhanced by the last two equations of (1.1).

Next, we prove the global existence of solutions to system (1.1) subject to initial condition (ic). Due to Theorem 3.2, we can reformulate the system (1.1) as

$$\begin{aligned} p_t^+ + sp_x^+ &= -\mu p^+ + \mu q^+(p^+ + p^-), \\ p_t^- - sp_x^- &= -\mu p^- + \mu q^-(p^+ + p^-), \\ \xi_t &= \kappa(p^+ - p^-)(1 - \xi^2), \end{aligned} \tag{3.6}$$

where q^+ and q^- are given by

$$q^+ = \frac{1+\xi}{2}, \qquad q^- = \frac{1-\xi}{2}. \tag{3.7}$$

It is worthwhile to note that here ξ represents the expectation of fiber orientation in one dimension subject to the initial condition $\xi_I := \xi(0) = q_I^+ - q_I^-$. Furthermore from initial condition (ic), we have

$$-1 \leq \xi_I \leq 1.$$

We seek the global solutions of the system (3.6) in the following space:

$$\mathbb{X}(0, T) := \{(p^+, p^-, \xi) |\ p^\pm, \xi \in L^\infty(0, T; L^1 \cap L^\infty(\mathbb{R}))\}.$$

We first give the local existence of solutions for the system (3.6).

Lemma 3.3 (local existence). *Let* $p_I^\pm, q_I^\pm(x) \geq 0$ *and* $q_I^+ + q_I^- = 1$. *Assume* $p_I^\pm \in L^1 \cap L^\infty(\mathbb{R})$ *and* $\xi_I \in L^1(\mathbb{R})$. *Then there exists a time* $T_0 > 0$ *such that the problem* (3.6) *with boundary condition* (bc1) *has a unique solution* $(p^+, p^-, \xi) \in \mathbb{X}(0, T_0)$ *satisfying* $-1 \leq \xi \leq 1$.

Proof. For short we denote $\eta = (p^+, p^-, \xi)^T$. The norm of the vector η is defined as

$$\begin{aligned} \|\eta\|_{L^\infty(\mathbb{R})} &= \max\{\|p^+\|_{L^\infty(\mathbb{R})}, \|p^-\|_{L^\infty(\mathbb{R})}, \|\xi\|_{L^\infty(\mathbb{R})}\}, \\ \|\eta\|_{L^1(\mathbb{R})} &= \max\{\|p^+\|_{L^1(\mathbb{R})}, \|p^-\|_{L^1(\mathbb{R})}, \|\xi\|_{L^1(\mathbb{R})}\}, \end{aligned}$$

Moreover, for the convenience of presentation we denote

$$\begin{aligned} f_1(p^+, p^-, \xi) &= -\mu p^+ + \frac{\mu}{2}(1 + \xi)(p^+ + p^-), \\ f_2(p^+, p^-, \xi) &= -\mu p^- + \frac{\mu}{2}(1 - \xi)(p^+ + p^-), \\ f_3(p^+, p^-, \xi) &= \kappa(p^+ - p^-)(1 - \xi^2). \end{aligned}$$

Clearly the function $f_i (i = 1, 2, 3)$ is differentiable with respect to its arguments and hence is locally Lipschitz continuous in any bounded subset of $L^1 \cap L^\infty(\mathbb{R})$.

It is straightforward to show that system (3.6) is strictly hyperbolic with three distinct uniform bounded eigenvalues λ_1, λ_2 satisfying $-s = \lambda_1 < \lambda_3 = 0 < \lambda_2 = s$. Then for each $i = 1, 2, 3$ and each point (t, x) in the $t - x$ plane, the characteristic equation of (3.6) defined by

$$\frac{d\mathbf{x}_i}{d\tau} = \lambda_i, \qquad \mathbf{x}_i(t) = x,$$

has a unique solution defined for all $t > 0$, describing the ith characteristic through point (t, x). We denote such a solution by $t \mapsto \mathbf{x}_i(\tau; t, x)$, where $\mathbf{x}_i(\tau; t, x) = x + \lambda_i(\tau - t)$ and in particular $\mathbf{x}_3(\tau; t, x) = x$ due to $\lambda_3 = 0$. Following the argument in [2], we define a set

$$\mathcal{D} = \left\{(t, x) \mid 0 \leq t < \ell/s, -\ell + st \leq x \leq l - st\right\}.$$

Note that ℓ can be arbitrarily large since the domain is unbounded. Then for every $(t, x) \in \mathcal{D}$ and every $i \in \{1, 2\}$, the characteristic curve $\{(t, x_i(\tau; t, x))| \ 0 \leq \tau \leq t\}$ is entirely contained inside $\mathcal{D}$ with $\mathbf{x}_i(0; t, x) \in [-\ell, \ell]$. Such a set $\mathcal{D}$ is called a domain of determinacy (see [2]).

The system (3.6) has two independent characteristics. We integrate the first equation of (3.6) along the second characteristic curve $\mathbf{x}_2(\tau; t, x)$, the second equation of (3.6) along the first characteristic $\mathbf{x}_1(\tau; t, x)$, and the third equation along $\mathbf{x}_3(\tau; t, x) = x$. Then (3.6) can be rewritten as an ODE system

$$\begin{aligned} p^+_\tau &= -\mu p^+(\tau, \mathbf{x}_2(\tau)) + \mu q^+(\tau, \mathbf{x}_2(\tau))(p^+(t, \mathbf{x}_2(\tau)) + p^-(\tau, \mathbf{x}_2(\tau))),\\ p^-_\tau &= -\mu p^-(\tau, \mathbf{x}_1(\tau)) + \mu q^-(\tau, \mathbf{x}_1(\tau))(p^+(\tau, \mathbf{x}_1(\tau)) + p^-(\tau, \mathbf{x}_1(\tau))),\\ \xi_\tau &= \kappa(p^+(\tau, x) - p^-(\tau, x))(1 - \xi^2(\tau, x)), \end{aligned} \tag{3.8}$$

where $\mathbf{x}_i(\tau) := \mathbf{x}_i(\tau; t, x)$ for $i = 1, 2$ and $\mathbf{x}_3(\tau) = x$.

In vector form, (3.8) can be reformulated as

$$u_\tau = f(u), \qquad u \in \mathbb{R}^3,$$

where

$$f(u) = \begin{pmatrix} f_1(u(\tau, \mathbf{x}_2(\tau))) \\ f_2(u(\tau, \mathbf{x}_1(\tau))) \\ f_3(u(\tau, x)) \end{pmatrix}.$$

Note that $\mathbf{x}_i(\tau) \in \mathbb{R}$ $(i = 1, 2)$. Then $f(u)$ is locally Lipschitz continuous in any bounded subset of $L^1 \cap L^\infty(\mathbb{R})$, and hence the local existence follows from the fundamental theorem of existence and uniqueness (e.g., see [22]). Due to Theorem 3.2 and the definition of ξ, we have that $-1 \leq \xi \leq 1$. Then the proof is finished. □

We proceed to derive a priori estimates in order to get global existence.

LEMMA 3.4 (a priori estimates). *Let the assumptions in Lemma* 3.3 *hold and let* (p^+, p^-, ξ) *be the solution obtained in Lemma* 3.3. *Then for any* $0 < t \leq T_0$, *there exist constants* $C > 0$ *and* $\widetilde{C} > 0$ *such that*

$$\|p^+(t)\|_{L^1 \cap L^\infty(\mathbb{R})} + \|p^-(t)\|_{L^1 \cap L^\infty(\mathbb{R})} + \|\xi(t)\|_{L^1 \cap L^\infty(\mathbb{R})} \leq C \exp(\widetilde{C} T),$$

and $-1 \leq \xi \leq 1$, *where* $\| \cdot \|_{L^1 \cap L^\infty(\mathbb{R})} = \| \cdot \|_{L^1(\mathbb{R})} + \| \cdot \|_{L^\infty(\mathbb{R})}$.

Proof. For each $(t,x) \in \mathcal{D}$ and $\mathbf{x}_i(0;t,x) \in [-\ell,\ell]$, we integrate the first two equations of (3.8) with respect to τ over $[0,t]$ and obtain that

$$\begin{aligned} &p^+(t,x) = p^+(\mathbf{x}_2(0)) + \int_0^t f_1\big(p^+(\tau,\mathbf{x}_2(\tau)), p^-(\tau,\mathbf{x}_2(\tau)), \xi(\tau,\mathbf{x}_2(\tau))\big)d\tau,\\ &p^-(t,x) = p^-(\mathbf{x}_1(0)) + \int_0^t f_2\big(p^+(\tau,\mathbf{x}_1(\tau)), p^-(\tau,\mathbf{x}_1(\tau)), \xi(\tau,\mathbf{x}_1(\tau))\big)d\tau,\\ &\xi(t,\xi) = \xi_I + \int_0^t (p^+(\tau,x) - p^-(\tau,x))(1-\xi^2(\tau,x))d\tau. \end{aligned} \tag{3.9}$$

Using the terminology from [2], we call (p^+,p^-,ξ) a *broad* solution for the Cauchy problem of (3.8) if (p^+,p^-,ξ) satisfies (3.9), at almost every point $(t,x)\in\mathcal{D}$. In the circumstance of semigroup theory, the broad solution defined above is called a mild solution if the transport operator in (3.6) generates a continuous semigroup (see [13] for details).

Taking the L^∞-norm on both sides of (3.9), using the fact that f_i is Lipschitz continuous, and taking into account $f_i(0,0,\xi)=0$ for $i=1,2$, we infer that

$$\begin{aligned} &\|p^+(t)\|_{L^\infty(\mathbb{R})} + \|p^-(t)\|_{L^\infty(\mathbb{R})} + \|\xi(t)\|_{L^\infty(\mathbb{R})}\\ &\quad \le C_1 + C_2\int_0^t (\|p^+(\tau)\|_{L^\infty(\mathbb{R})} + \|p^-(\tau)\|_{L^\infty(\mathbb{R})} + \|\xi(\tau)\|_{L^\infty(\mathbb{R})})d\tau, \end{aligned}$$

where C_1 is a constant such that $\|p_I^+\|_{L^\infty(\mathbb{R})} + \|p_I^-\|_{L^\infty(\mathbb{R})} + \|\xi_I\|_{L^\infty(\mathbb{R})} \le C_1$ and C_2 depends on the Lipschitz constants of the functions $f_i (i=1,2,3)$ and the turning rate μ.

The application of Gronwall's inequality to the above inequality gives

$$\|p^+(t)\|_{L^\infty(\mathbb{R})} + \|p^-(t)\|_{L^\infty(\mathbb{R})} + \|\xi(t)\|_{L^\infty(\mathbb{R})} \le C_1\exp(C_2 t).$$

Similarly, one can deduce that there exist constants $C_3, C_4 > 0$ such that

$$\|p^+(t)\|_{L^1(\mathbb{R})} + \|p^-(t)\|_{L^1(\mathbb{R})} + \|\xi(t)\|_{L^1(\mathbb{R})} \le C_3\exp(C_4 t).$$

The last two inequalities imply the first conclusion of the lemma. The second conclusion $-1\le\xi\le 1$ follows directly from Theorem 3.2 and the definition of ξ. ☐

By Lemmas 3.3 and 3.4, the existence theorem of global solutions is obtained.

THEOREM 3.5 (global existence). *Let initial condition* (ic) *hold. Assume* $p_I^\pm, \xi_I \in L^1\cap L^\infty(\mathbb{R})$. *Then the problem* (3.6) *with boundary condition* (bc1) *has a unique global solution* $(p^+,p^-,\xi)\in\mathbb{X}(0,\infty)$ *satisfying* $-1\le\xi\le 1$ *and* $p^\pm \ge 0$. *Consequently, the problem* (1.1) *with initial condition* (ic) *and boundary condition* (bc1) *has a unique global solution* (p^+,p^-,q^+,q^-) *such that* $p^\pm, q^\pm \in L^\infty(0,\infty; L^1\cap L^\infty(\mathbb{R}))$ *with* $p^\pm\ge 0$ *and* $0\le q^\pm\le 1$ *with* $q^+ + q^- = 1$.

Proof. We suppose that the maximal time $T_{\max}$ of existence for the solution of (3.6) is finite, namely, $T_{\max}<\infty$. From Lemma 3.4, we know that $-1\le\xi\le 1$ for any $0\le t\le T_{\max}$. Hence according to the well-known alternative results (e.g., see [20, 22]), one has that

$$\lim_{t\to T_{\max}} \|p^+(t)\|_{L^1\cap L^\infty(\mathbb{R})} = \infty \quad \text{or} \quad \lim_{t\to T_{\max}} \|p^-(t)\|_{L^1\cap L^\infty(\mathbb{R})} = \infty. \tag{3.10}$$

On the other hand, when $-1 \leq \xi \leq 1$, we have proven in Lemma 3.4 that for any $t \leq T_{\max}$, it holds that

$$\|p^+(t)\|_{L^1\cap L^\infty(\mathbb{R})} + \|p^-(t)\|_{L^1\cap L^\infty(\mathbb{R})} \leq C\exp(\widetilde{C}T_{\max}),$$

which contradicts (3.10) for $0 < T_{\max} < \infty$. This contradiction in turn shows that $T_{\max} = \infty$, and hence the global solution of (3.6) follows. Due to Theorem 3.1, the second conclusion is an immediate consequence. $\square$

Remark 2. Mathematically, when cutting efficiency $\kappa = 0$, the system (1.1) becomes the one-dimensional mesenchymal motion model for undirected tissue (see [12]). Due to the assumption $q^+(t,x) = q^-(t,x)$ for undirected tissue, we obtain the following global existence theorem for the model associated with undirected tissue.

THEOREM 3.6. *Suppose $\kappa = 0$. Let initial condition* (ic) *hold and let $q_I^+ = q_I^- = 1/2$. Assume $p_I^\pm \in L^1 \cap L^\infty(\mathbb{R})$. Then there exists a unique global solution to system* (3.6) *such that $(p^+, p^-, \xi) \in \mathbb{X}(0,\infty)$ with $\xi = 0$ and $p^\pm \geq 0$. Hence there is a unique global solution $(p^+, p^-, 1/2, 1/2)$ to* (1.1) *with initial condition* (ic) *and boundary condition* (bc1) *such that $p^\pm \in L^\infty(0,\infty; L^1 \cap L^\infty(\mathbb{R}))$ satisfying $p^\pm \geq 0$.*

Since the functions on the right-hand side of (1.1) are continuously differentiable with respect to p^+, p^-, q^+, and q^-, by a theory for semilinear hyperbolic systems in [2] (see Theorem 3.6 in [2]), the broad solution of Cauchy problem (1.1) obtained in Theorem 3.5 is indeed a classical solution provided that the initial data (1.9) are continuously differentiable, namely, we have the following results.

THEOREM 3.7. *Let the assumptions in Theorem* 3.5 *hold. In addition, we assume that the initial data in* (1.9) *are continuously differentiable. Then the broad solution $u : \mathcal{D} \to \mathbb{R}^2$ obtained in Theorem* 3.5 *provides a classical solution. Moreover, if initial data in* (1.9) *are nonnegative, the solution is nonnegative. Its partial derivatives u_t and u_x, respectively, are broad solutions of the following semilinear system:*

$$\begin{aligned}(u_t)_t &= H_u u_t - \Theta \cdot (u_t)_x,\\ (u_x)_t &= H_u u_x - \Theta \cdot (u_x)_x,\end{aligned}$$

where u, H, and Θ are defined as in section 2 *and H_u denotes the derivative of H with respect to u.*

Proof. The proof is similar to the argument in [2]. We omit the details. $\square$

4. Macroscopic limits. For the given fiber distribution $q^\pm(t,x)$, formal parabolic and hydrodynamic limits were derived in [12] for the mesenchymal motion models (1.1) in $n(n \geq 1)$ dimensions. In this section we rigorously carry out the parabolic limits for system (1.1) under some suitable assumptions.

To derive a limiting diffusion model for (1.1), we use the parabolic scaling of space and time, with $\bar{x} = \varepsilon x$ denoting a macroscopic space scale and $\bar{t} = \varepsilon^2 t$ a long time scale. Now we use the equivalent system (1.3) in a slightly different form using the flux $J = p^+ - p^-$. Upon substituting the above scaling variable into (1.3), and dropping the bar for convenience, we end up with the following equations:

$$\begin{aligned}\varepsilon^2\partial_t p_\varepsilon + \varepsilon s \partial_x J_\varepsilon &= 0,\\ \varepsilon^2\partial_t J_\varepsilon + \varepsilon s \partial_x p_\varepsilon &= \mu\xi_\varepsilon p_\varepsilon - \mu J_\varepsilon,\\ \varepsilon^2\partial_t \xi_\varepsilon &= \kappa(p_\varepsilon^+ - p_\varepsilon^-)(1-\xi_\varepsilon^2),\end{aligned} \tag{4.1}$$

with initial data $p_\varepsilon(0) = p_I = p_I^+ + p_I^-$, $J_\varepsilon(0) = J_I = p_I^+ - p_I^-$, $\xi_\varepsilon(0, \cdot) = q_I^+ - q_I^-$. The system (4.1) is equivalent to the following second-order damped hyperbolic equation (see (1.5) or [12]):

$$\frac{\varepsilon^4}{\mu}\partial_t^2 p_\varepsilon + \varepsilon^2 \partial_t p_\varepsilon + \varepsilon \partial_x (s\xi_\varepsilon p_\varepsilon) = \varepsilon^2 \frac{s^2}{\mu}\partial_x^2 p_\varepsilon, \tag{4.2}$$

which indicates that the drift term is a dominating term for ε small. As in [12], we assume that the expectation of fiber directions is small as to the order of ε:

$$\xi_q(t,x) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\xi_\varepsilon\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[q^+\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) - q^-\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right)\right] < \infty. \tag{4.3}$$

Under the above assumption, we formally obtain a drift-diffusion model with diffusion coefficient $\frac{s^2}{\mu}$ and drift velocity $s\xi_q$ from (4.2) by sending $\varepsilon \to 0$ (see [12]),

$$\partial_t p + \partial_x(s\xi_q p) = \frac{s^2}{\mu}\partial_x^2 p, \tag{4.4}$$

where p is the limit of p_ε as $\varepsilon \to 0$. The goal of this section is to show that the solution of (4.2) is convergent to the solution of (4.4) in the weak sense as $\varepsilon \to 0$. To proceed we give the definition of weak solutions that we address here.

DEFINITION 4.1. *We say that a function $P \in L^\infty([0,T]; H^1(\mathbb{R}))$ is a weak solution of* (4.4) *if $P(t,x)$ satisfies the following:*

(a) *For any test function $\phi \in C_0^\infty([0,T) \times \mathbb{R})$, it holds that*

$$-\int_0^T \int_{\mathbb{R}} P\partial_t\phi dxdt - \int_0^T\int_{\mathbb{R}} (s\xi_q P)\partial_x\phi dxdt = \frac{s^2}{\mu}\int_0^T\int_{\mathbb{R}} P\partial_x^2\phi dxdt + \int_{\mathbb{R}} P(0)\phi(0)dx.$$

(b) $P(0) = p_I = p_I^+ + p_I^-$.

Next we establish the convergence properties of the solution $(p_\varepsilon,\ J_\varepsilon)$ as $\varepsilon \to 0$. It suffices to derive a uniform estimate for the solutions of system (4.1), which is given in the following lemma.

LEMMA 4.2. *Let $p_I^\pm \in H^1(\mathbb{R})$ and let the assumption* (4.3) *hold. Assume further that there exists a constant $C_1 > 0$, independent of ε, such that*

$$|\xi_\varepsilon|, |\partial_x\xi_\varepsilon| \le C_1\varepsilon. \tag{4.5}$$

Then there is a constant C_2, independent of ε, such that the solution $(p_\varepsilon, J_\varepsilon)$ of system (4.1) *satisfies, for any $0 \le t \le T$,*

$$\begin{aligned} &\|p_\varepsilon(t)\|_{H^1(\mathbb{R})} + \|J_\varepsilon(t)\|_{H^1(\mathbb{R})} + \|\varepsilon\partial_t p_\varepsilon\|_{L^2(\mathbb{R})} \\ &\quad \le C_2(C_1,\mu,T)(\|p_I\|_{H^1(\mathbb{R})} + \|J_I\|_{H^1(\mathbb{R})}), \end{aligned} \tag{4.6}$$

where the constant C_2 depends on C_1, μ, and T.

Proof. We use the energy method to prove the lemma. First, note that $p_\varepsilon(0) = p_I = p_I^+ + p_I^- \in H^1(\mathbb{R})$ and $J_\varepsilon(0) = J_I = p_I^+ - p_I^- \in H^1(\mathbb{R})$. Multiplying the first equation of (4.1) by p_ε and the second by J_ε, adding the resultant equations, and integrating over $[0,t) \times \mathbb{R}$, we end up with the following inequality:

$$\begin{aligned}
&\frac{1}{2}\int_{\mathbb{R}}(|p_\varepsilon|^2 + |J_\varepsilon|^2)dx + \int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}|J_\varepsilon|^2 dxd\tau \\
&\quad = \frac{1}{2}\int_{\mathbb{R}}(|p_I|^2 + |J_I|^2)dx + \int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}\xi_\varepsilon p_\varepsilon J_\varepsilon dxd\tau \\
&\quad \leq \frac{1}{2}\int_{\mathbb{R}}(|p_I|^2 + |J_I|^2)dx + \int_0^t \int_{\mathbb{R}} \mu C_1|\varepsilon^{-1}p_\varepsilon J_\varepsilon|dxd\tau,
\end{aligned} \tag{4.7}$$

where we have used the assumption (4.5). Applying Young's inequality $|C_1\varepsilon^{-1}p_\varepsilon J_\varepsilon| \leq \frac{1}{2}(\varepsilon^{-2}|J_\varepsilon|^2 + C_1^2|p_\varepsilon|^2)$ in (4.7), we have

$$\begin{aligned}
&\int_{\mathbb{R}}(|p_\varepsilon|^2 + |J_\varepsilon|^2)dx + \int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}|J_\varepsilon|^2 dxd\tau \\
&\qquad \leq \int_{\mathbb{R}}(|p_I|^2 + |J_I|^2)dx + \mu C_1^2 \int_0^t \int_{\mathbb{R}} |p_\varepsilon|^2 dxd\tau.
\end{aligned}$$

By Gronwall's inequality, we immediately get an L^2-estimate of p_ε and J_ε independent of ε such that for $0 \leq t < T$,

$$\|p_\varepsilon\|^2_{L^2(\mathbb{R})} + \|J_\varepsilon\|^2_{L^2(\mathbb{R})} \leq (\|p_I\|^2_{L^2(\mathbb{R})} + \|J_I\|^2_{L^2(\mathbb{R})})\exp(\mu C_1^2 T). \tag{4.8}$$

Next we go to the higher order estimates. To this end, we multiply the first equation of (4.1) by $-\partial_x^2 p_\varepsilon$ and the second by $-\partial_x^2 J_\varepsilon$. Then we end up with the following estimates using the same procedure as that deriving (4.7):

$$\begin{aligned}
&\frac{1}{2}\int_{\mathbb{R}}(|\partial_x p_\varepsilon|^2 + |\partial_x J_\varepsilon|^2)dx + \int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}|\partial_x J_\varepsilon|^2 dxd\tau \\
&= \frac{1}{2}\int_{\mathbb{R}}(|\partial_x p_I|^2 + |\partial_x J_I|^2)dx + \int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}\partial_x(\xi_\varepsilon p_\varepsilon)\partial_x J_\varepsilon dxd\tau \\
&\leq \frac{1}{2}\int_{\mathbb{R}}(|\partial_x p_I|^2 + |\partial_x J_I|^2)dx + \int_0^t \int_{\mathbb{R}} \mu C_1\varepsilon^{-1}(|p_\varepsilon| + |\partial_x p_\varepsilon|)|\partial_x J_\varepsilon|dxd\tau.
\end{aligned}$$

Using Young's inequality and the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$, we deduce that

$$\begin{aligned}
&\int_0^t \int_{\mathbb{R}} \mu C_1\varepsilon^{-1}(|p_\varepsilon| + \partial_x p_\varepsilon|)|\partial_x J_\varepsilon|dxd\tau \\
&\leq \frac{1}{2}\int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}|\partial_x J_\varepsilon|^2 dxd\tau + \frac{C_1^2}{2}\int_0^t \int_{\mathbb{R}} \mu(|p_\varepsilon| + |\partial_x p_\varepsilon|)^2 dxd\tau \\
&\leq \frac{1}{2}\int_0^t \int_{\mathbb{R}} \mu\varepsilon^{-2}|\partial_x J_\varepsilon|^2 dxd\tau + C_1^2\int_0^t \int_{\mathbb{R}} \mu|\partial_x p_\varepsilon|^2 dxd\tau + C(T, p_I, J_I),
\end{aligned} \tag{4.9}$$

where (4.8) has been used and

$$C(T, p_I, J_I) = \mu C_1^2 T(\|p_I\|^2_{L^2(\mathbb{R})} + \|J_I\|^2_{L^2(\mathbb{R})})\exp(\mu C_1^2 T).$$

Now substituting (4.9) into (4.7) and applying Gronwall's inequality to the resulting inequality, we infer that

$$\begin{aligned}
&\|\partial_x p_\varepsilon\|^2_{L^2(\mathbb{R})} + \|\partial_x J_\varepsilon\|^2_{L^2(\mathbb{R})} \\
&\quad \leq C(T, p_I, J_I)(\|\partial_x p_I\|^2_{L^2(\mathbb{R})} + \|\partial_x J_I\|^2_{L^2(\mathbb{R})})\exp(\mu C_1^2 T) \\
&\quad \leq \mu C_1^2 T(\|p_I\|_{H^1(\mathbb{R})} + \|J_I\|_{H^1(\mathbb{R})})^2 \exp(2\mu C_1^2 T).
\end{aligned} \tag{4.10}$$

Furthermore, by (4.1) we have

$$\|\varepsilon\partial_t p_\varepsilon\|_{L^2(\mathbb{R})} = \|\partial_x J_\varepsilon\|_{L^2(\mathbb{R})}. \tag{4.11}$$

Then the combination of (4.8), (4.10), and (4.11) gives (4.6) and completes the proof. □

THEOREM 4.3. *Let the assumptions in Lemma* 4.2 *hold and let* $p_\varepsilon(0) = p_I = p_I^+ + p_I^-$. *Then as* $\varepsilon \to 0$, *the solutions* p_ε *of* (4.2) *converge to a limit function* p_0, *which is a weak solution of* (4.4) *such that* $p_0(t = 0) = p_I$.

Proof. According to the energy estimates (4.6), we see that the solution sequence p_ε is uniformly bounded in $L^\infty_{\text{loc}}([0,\infty); H^1(\mathbb{R}))$ and $\varepsilon\partial_t p_\varepsilon$ is uniformly bounded in $L^\infty_{\text{loc}}([0,\infty); L^2(\mathbb{R}))$ for every $\varepsilon > 0$.

As a consequence of the Rellich–Kondrachov compactness theorem, there exist a subsequence of p_ε and $\varepsilon\partial_t p_\varepsilon$, still denoted by p_ε and $\varepsilon\partial_t p_\varepsilon$, and functions $p_0 \in L^\infty_{\text{loc}}([0,\infty); H^2(\mathbb{R}))$ and $p_1 \in L^\infty_{\text{loc}}([0,\infty); L^2(\mathbb{R}))$ such that

$$\begin{cases} p_\varepsilon \rightharpoonup p_0 \quad \text{weakly}^* \text{ in } \ L^\infty_{\text{loc}}([0,\infty); H^1(\mathbb{R})), \\ \varepsilon\partial_t p_\varepsilon \rightharpoonup p_1 \quad \text{weakly}^* \text{ in } \ L^\infty_{\text{loc}}([0,\infty); L^2(\mathbb{R})). \end{cases} \tag{4.12}$$

Next we show that p_0 is a weak solution of (4.4) subject to the given initial data. To this end we multiply (4.2) by a test function $\phi \in C_0^\infty([0,T) \times \mathbb{R})$ with $\phi(T) = 0$ and integrate the resultant equation to get

(4.13)
$$\begin{aligned} &\frac{\varepsilon^2}{\mu}\int_0^T\int_{\mathbb{R}} p_\varepsilon \partial_t^2\phi dxdt + \frac{\varepsilon^2}{\mu}\int_{\mathbb{R}}[p_\varepsilon(T)\partial_t\phi(T) - \partial_t p_\varepsilon(0)\phi(0)]dx \\ &\quad - \frac{\varepsilon^2}{\mu}\int_{\mathbb{R}}[\partial_t p_\varepsilon(T)\phi(T) - p_\varepsilon(0)\partial_t\phi(0)]dx - \int_0^T\int_{\mathbb{R}} p_\varepsilon\partial_t\phi dxdt + \int_{\mathbb{R}} p_\varepsilon(T)\phi(T)dx \\ &\quad - \frac{1}{\varepsilon}\int_0^T\int_{\mathbb{R}}(s\xi_\varepsilon p_\varepsilon)\partial_x\phi dxdt = \int_{\mathbb{R}} p_\varepsilon(0)\phi(0)dx + \frac{s^2}{\mu}\int_0^T\int_{\mathbb{R}} p_\varepsilon\partial_x^2\phi dxdt. \end{aligned}$$

Note that $p_\varepsilon(0) = p_I = p_I^+ + p_I^- \in H^1(\mathbb{R})$. Hence $J_\varepsilon(0) = J_I = p_I^+ - p_I^- \in H^1(\mathbb{R})$ and $\varepsilon\partial_t p_\varepsilon(0) = \partial_x J_\varepsilon(0) \in L^2(\mathbb{R})$ from (4.1). Thus the second, third, and fourth terms in (4.13) vanish as $\varepsilon \to 0$ by (4.12). Using assumption (4.3) and sending $\varepsilon \to 0$ in (4.13), we obtain from (4.12) that

$$\begin{aligned} &-\int_0^T\int_{\mathbb{R}} p_0\partial_t\phi dxdt - \int_0^T\int_{\mathbb{R}}(s\xi_q p_0)\partial_x\phi dxdt \\ &\quad = \int_{\mathbb{R}} p_I\phi(0)dx + \frac{s^2}{\mu}\int_0^T\int_{\mathbb{R}} p_0\partial_x^2\phi dxdt, \end{aligned} \tag{4.14}$$

which shows that p_0 is a weak solution of (4.4) satisfying the initial condition. □

Remark 3. It is worthwhile to note that assumptions (4.5) and (4.3) are automatically satisfied for the case of undirected tissue where $\xi_\varepsilon = 0$ (see also Remark 2). Then the limit equation for the case of undirected tissue is a pure diffusion equation without a drift term.

5. Traveling waves. Since the system (1.1) models the invasion of cells through tissue, it is of interest to look for traveling wave solutions for (1.1) and see what kinds of movement patterns are used by cells for invasion. To this end, we first use the invariant of motion $q^+ + q^- = 1$ and consider the equivalent system (1.4).

We introduce the wave variable

$$z = x - ct,$$

where $c \geq 0$ denotes the wave speed. Then we can define the wave profile by

$$\begin{aligned} p(z) &= p(t,x) = p(x-ct), \\ j(z) &= j(t,x) = j(x-ct), \\ q^+(z) &= q^+(t,x) = q^+(x-ct). \end{aligned} \tag{5.1}$$

Substituting (5.1) into (1.4), we convert (1.4) into an ODE system as follows:

$$\begin{aligned} -cp_z + j_z &= 0, \\ -cj_z + s^2 p_z &= -\mu j + \mu s(2q^+ - 1)p, \\ -cq_z^+ &= \frac{2\kappa}{s} j(1-q^+)q^+. \end{aligned} \tag{5.2}$$

We prescribe the boundary conditions by

$$\begin{aligned} p(-\infty) &= p(+\infty) = 0, \\ j(-\infty) &= j(+\infty) = 0, \\ q^+(-\infty) &= q_l^+,\ q^+(+\infty) = q_r^+, \end{aligned} \tag{5.3}$$

where q_l^- and q_r^+ are constants and satisfy $0 \leq q_l^-,\, q_r^+ \leq 1$, and $q_l^- > q_r^+$. That is, we look for the traveling pulse wave for p and decreasing traveling front wave for q^+.

From (5.2) and the boundary conditions (5.3), we obtain an invariant of motion for j and p such that

$$j = cp. \tag{5.4}$$

Then the system (5.2) is reduced to a two-dimensional system by the substitution of (5.4) into (5.2):

$$\begin{aligned} (c^2 - s^2)p_z &= \mu p[c - s(2q^+ - 1)], \\ q_z^+ &= -\frac{2\kappa}{s} p(1-q^+)q^+. \end{aligned} \tag{5.5}$$

It is clear that (5.5) becomes a singular problem when $c = s$ and that this singular problem has no solution satisfying the boundary conditions (5.3). Indeed if $c = s$, then $q^+ = 1$ due to $\mu \neq 0$, which biologically means cells continuously move to the right without changing movement direction. Also, $q^+ = 1$ does not agree with the boundary conditions (5.3). Thus we assume $c \neq s$ hereafter. We will see later that biologically meaningful waves exist only for $c < s$. However, for now, we just assume $c \neq s$, and system (5.5) can be rewritten as

$$\begin{aligned} p_z &= -\alpha p[c - s(2q^+ - 1)], \\ q_z^+ &= -\beta p(1-q^+)q^+, \end{aligned} \tag{5.6}$$

where $\alpha = -\frac{\mu}{c^2 - s^2}$, $\beta = \frac{2\kappa}{s} > 0$. Due to the biological interest, we consider only nonnegative solutions where $p \geq 0$ and $0 \leq q^\pm \leq 1$. In fact, the nonnegativity of solutions to the system (5.6) with boundary conditions (5.3) can be analogously obtained by following the argument used in section 3. Therefore we are interested only in those heteroclinic orbits that remain nonnegative.

5.1. Phase plane analysis. System (5.6) has a continuum of steady states $(0,\theta)$ with $0 \le \theta \le 1$. The Jacobian matrix linearized about the steady state $(0,\theta)$ is

$$J_s = \begin{bmatrix} -\alpha\big(c - s(2\theta - 1)\big) & 0 \\ -\beta(1-\theta)\theta & 0 \end{bmatrix}.$$

The eigenvalues of J_s are

$$\lambda_1 = -\alpha\big(c - s(2\theta - 1)\big), \qquad \lambda_2 = 0. \tag{5.7}$$

The corresponding eigenvectors are

$$r_1 = \begin{bmatrix} \lambda_1 \\ -\beta(1-\theta)\theta \end{bmatrix}, \qquad r_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{5.8}$$

When $c \ne s$, we have two cases to consider corresponding to the sign of eigenvalue λ_1.

Case 1. If $c > s > 0$, then $\alpha < 0$. It is straightforward to check that $\lambda_1 > 0$, which indicates every steady state $(0,\theta)$ with $0 \le \theta \le 1$ is unstable, and consequently there is no nonnegative heteroclinic connection due to the lack of the stable manifold. We thus claim that $0 \le c < s$ is a necessary condition for the existence of a traveling wave and s is then a critical traveling speed. Thus we assume that $c < s$ hereafter.

Case 2. If $0 \le c < s$, then $\alpha > 0$. We first fix the traveling speed c and solve $c - s(2\theta^* - 1) = 0$ to get $\theta^* = \frac{c+s}{2s}$. Clearly we have that $0 < \theta^* < 1$. Furthermore the following properties hold:

$$\begin{aligned} &\theta < \theta^* \Rightarrow \lambda_1 < 0, \\ &\theta = \theta^* \Rightarrow \lambda_1 = 0, \\ &\theta > \theta^* \Rightarrow \lambda_1 > 0. \end{aligned} \tag{5.9}$$

Next, we show that there exists a pair of equilibria which generates a heteroclinic connection for each fixed c satisfying $0 \le c < s$. From (5.7), we see that every steady state $(0,\theta)$ of the system (5.6) with $0 \le \theta \le 1$ has two manifolds, one of which is a one-dimensional center manifold corresponding to zero eigenvalue λ_2. Since each center manifold is invariant under the flow of the system (5.6), and the set $\{(p,q^+) : p = 0, 0 \le q^+ \le 1\}$ consists of steady states only and hence is invariant, the center manifold is the q^+ axis where $0 \le q^+ \le 1$. So the heteroclinic connection is determined only by the stable and unstable manifolds corresponding to positive and negative eigenvalues given by λ_1, respectively. The existence of a heteroclinic orbit connecting the unstable manifold of one fixed point with the stable manifold of another fixed point corresponds to the existence of a traveling wave (heteroclinic orbit). Below we rigorously prove the existence of such a heteroclinic connection. Beyond this, we also shall prove the existence of a family of traveling waves since a continuum of steady state exists for the system (5.6). Before proceeding, we give a remark as follows.

Remark 4. The constants $q^+ = 0$ and $q^+ = 1$ are solutions of the second equation of (5.6), and furthermore it holds that

(a) if $q^+ = 0$, then $p \to +\infty$ as $z \to -\infty$;

(b) if $q^+ = 1$, then $p \to +\infty$ as $z \to +\infty$.

Therefore, neither the orbit $q^+ = 0$ nor $q^+ = 1$ can form a heteroclinic connection, although $\{q^+ = 1\}$ is the unstable manifold of the equilibrium $(0,1)$ and $\{q^+ = 0\}$ is the stable manifold of the equilibrium $(0,0)$. So hereafter we assume that $0 < q^+ < 1$ in order to obtain the existence of traveling waves.

5.2. Existence of traveling waves. To show that an unstable manifold can be connected by a stable manifold, we need to investigate the global structure of the original nonlinear system. Below we shall apply LaSalle's invariant principle (see [10, 18]) to study the asymptotic behavior of solutions of the system (5.6), which is described in the following lemma.

LEMMA 5.1. *Assume* $0 \le c < s$. *Let* (p, q^+) *be a solution of* (5.6) *subject to initial conditions* $p_I > 0$ *and* $0 < q_I^+ < 1$. *Then the* ω*-limit set of solutions to system* (5.6) *is contained in the following set:*

$$\mathbb{N} = \{(p, q^+) |\ p = 0,\ 0 < q^+ < \theta^*\}, \tag{5.10}$$

and the α*-limit set is contained in the set*

$$\mathbb{G} = \{(p, q^+) |\ p = 0,\ \theta^* < q^+ < 1\}, \tag{5.11}$$

where θ^* *is a constant between* 0 *and* 1 *determined by* $\theta^* = \frac{c+s}{2s}$.

Proof. Define a function $V(p, q^+)$ by $V(p, q^+) = q^+$. Then in the set $\{(p, q^+) |\ p \geq 0, 0 < q^+ < 1\}$, $V(p(z), q^+(z)) > 0$ and $\frac{dV}{dz} \le 0$ thanks to the second equation of (5.6). Given a number $L > 0$, we now define a set

$$\Omega_L = \{(p, q^+) |\ V(p, q^+) \le L, p > 0, 0 < q^+ < 1\}.$$

Since we restrict our attention to the case of $0 < q^+ < 1$, we let $0 < L < 1$. Hence it holds that

$$\Omega_L = \{(p, q^+) |\ p > 0, 0 < q^+ < L\}.$$

We now proceed to justify that the set Ω_L is bounded for given $0 < L < 1$. Toward this end, we divide the first equation of (5.6) by the second equation to obtain that

$$\frac{dp}{dq^+} = -\frac{\alpha(c+s)}{\beta}\frac{1}{(1-q^+)q^+} + \frac{2\alpha s}{\beta}\frac{1}{1-q^+}. \tag{5.12}$$

Integrating this equation and recovering α and β yield a first integral

$$p(q^+) = \frac{\mu s}{2\kappa}\left[\frac{\ln(1-q^+)}{c+s} - \frac{\ln q^+}{c-s}\right] + C, \tag{5.13}$$

where C is a constant of integration determined by the boundary condition of q^+ given in (5.3).

Then for any $q^+ = V(p, q^+) < L$, it is clear from (5.13) that p is bounded as a function of q^+. As a result, the set Ω_L defined above is bounded.

We now define another set

$$\mathbb{N}_1 = \left\{(p, q^+) \middle| \frac{dV}{dz} = 0, 0 < q^+ < 1\right\}.$$

From the second equation of (5.6), we know that

$$\frac{dV}{dz} = 0 \Longleftrightarrow p = 0 \text{ or } q^+ = 0 \text{ or } q^+ = 1.$$

Therefore, $\mathbb{N}_1 = \{(p, q^+) |\ p = 0, 0 < q^+ < 1\}$ and is invariant since it is composed of only steady states. With the help of LaSalle's invariant principle, the ω-limits set of

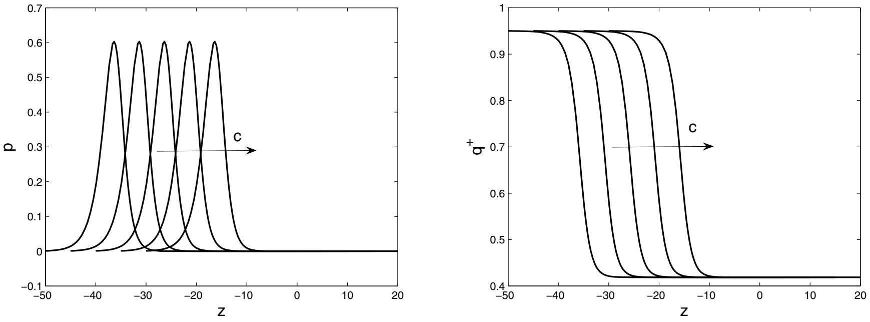


FIG. 1. *The traveling wave for the system* (5.6)*, where* $c = 1, s = 2, \mu = 2, \kappa = 1$*. The waves travel from the left to the right and* c *denotes the traveling speed and* $z = 0, 5, 10, 15, 20$*.*

any trajectories of the system starting in the set Ω_L for $0 < L < 1$ is contained in the set $\mathbb{N}_1$. Indeed, we can characterize the asymptotic behavior of the solution more precisely. From (5.9), we know that $\lambda_1 > 0$ for all $\theta^* < \theta < 1$. Then the equilibrium $(0, \theta)$ with $\theta^* < \theta < 1$ is unstable. If we define $\mathbb{N}_2 = \{(p, q^+)|\ p = 0, \theta^* < q^+ < 1\}$, then all solutions of the system (5.6) converge to the set as $z \to +\infty$:

$$\mathbb{N} = \mathbb{N}_1 \setminus \mathbb{N}_2 = \{(p, q^+)|\ p = 0,\ 0 < q^+ < \theta^*\}.$$

In a similar fashion, if we study the problem (5.6) backward on variable z, we can prove that all solutions of (5.6) converge to the set $\mathbb{G}$ when $z \to -\infty$, which completes the proof. □

Lemma 5.1 shows that any trajectory of the system (5.6) starting in a neighborhood of an equilibrium $(0, \theta)$ with $\theta^* < \theta < 1$ converges, as $z \to +\infty$, to another equilibrium $(0, \theta)$ with $0 < \theta < \theta^*$, which gives a nonnegative heteroclinic orbit (traveling wave) connecting these two equilibria. This heteroclinic orbit can be explicitly given by a level curve equation in the form of (5.13). It is worthwhile to point out that the traveling speed c can be 0 from our analysis, which corresponds to a standing wave. Hence we obtain the following existence theorem of traveling waves.

THEOREM 5.2. *Let us consider the system* (5.6) *given traveling speed* c *with* $0 \leq c < s$ *and* $\theta^* = \frac{c+s}{2s}$*. Then for any equilibrium* $(0, c_1)$ *with* $\theta^* < c_1 < 1$ *there exists another equilibrium* $(0, c_2)$ *with* $0 < c_2 < \theta^*$ *such that there is a bounded, nonnegative, heteroclinic orbit connecting* $(0, c_1)$ *to* $(0, c_2)$*. That is, there exists a traveling solution* (p, q^+) *of the system* (5.6) *connecting two equilibria. Particularly, the system* (5.6) *admits a standing wave for* $c = 0$*.*

Notice that in Lemma 5.3 we will give an explicit relation between c_1 and c_2.

An example of traveling solution (p, q^+) for system (5.6) is numerically plotted in Figure 1. From the definition of p and the relation (5.4), we can derive that

$$p^+ = \frac{s+c}{2s} p, \qquad p^- = \frac{s-c}{2s} p. \tag{5.14}$$

In addition to the relation

$$q^- = 1 - q^+, \qquad j = cp, \tag{5.15}$$

we find the traveling waves for p^+, p^-, q^-, and j in terms of p and q^+, as given above. The plots of the traveling structures of these quantities are given in Figure 2.

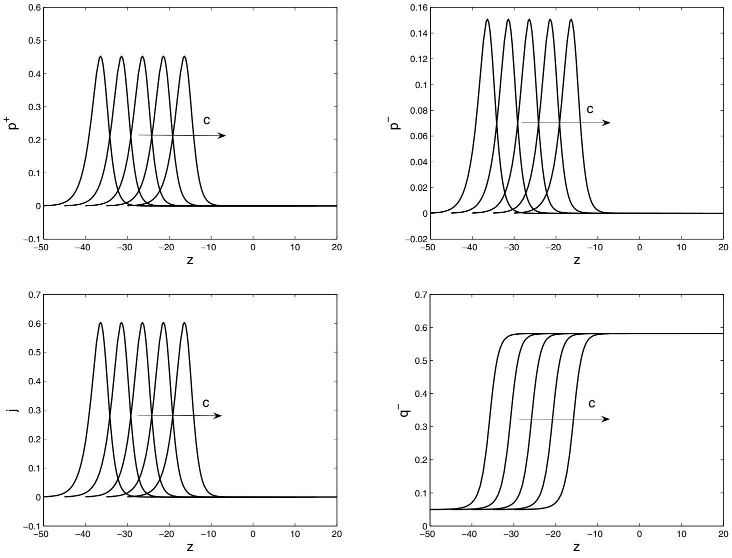


FIG. 2. *Numerical illustration of traveling waves for* p^+, p^-, j, *and* q^-, *where* $c = 1, s = 2, \mu = 2, \kappa = 1$. *The waves shift from the left to the right and* c *denotes the traveling speed and* $z = 0, 5, 10, 15, 20$.

A plot of all these quantities in a coordinate system is given in Figure 3 from which the transition properties between cell movement direction and fiber orientation are clearly indicated.

From the first equation of (1.4), we know that the total mass of cells is conserved and so traveling pulse waves are expected, as we found analytically and numerically above. The numerical simulation for p in Figure 1 indicates that individual cells can move to the left or the right, but the whole cell group will move to the right continuously. However, when the waves travel through, the fiber orientations are modified by cells, and alignment to cell movement direction is enhanced, which is indicated by the numerical simulation for q^+ in Figure 3.

5.3. Family of traveling waves. Note that for each left state q_l^+ with $\theta^* < q_l^+ < 1$ we find a corresponding right state $(0, q_r^+)$ connecting to $(0, q_l^+)$ which gives a traveling wave. Here we give an explicit formula which relates q_l^+ and q_r^+.

LEMMA 5.3. *Given a speed* c *satisfying* $0 \leq c < s$, *the left and right equilibria* $(0, q_l^+)$ *and* $(0, q_r^+)$ *are related as*

$$\left(\frac{1-q_r^+}{1-q_l^+}\right)^{s-c} = \left(\frac{q_l^+}{q_r^+}\right)^{s+c}, \qquad 0 \leq c < s. \tag{5.16}$$

Proof. An explicit heteroclinic connection has been given by (5.13). By Lemma 5.1, we infer that $p(q_l^+) = p(q_r^+) = 0$. Applying this condition to (5.13), one has that

$$\frac{\ln(1-q_l^+)}{c+s} - \frac{\ln q_l^+}{c-s} = \frac{\ln(1-q_r^+)}{c+s} - \frac{\ln q_r^+}{c-s}.$$

Rearranging the above identity yields (5.16). ▯

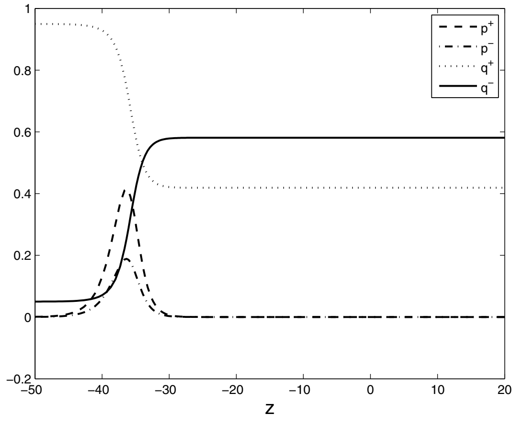


FIG. 3. *A plot of traveling solutions of system* (1.1) *in a coordinate system, where* $c = 1, s = 2, \mu = 2, \kappa = 1,$ *and* $z = 0, 5, 10, 15, 20$.

By Lemma 5.3 we identify a family of heteroclinic orbits as shown in Figure 4.

From (5.13) we see that p is bounded as a function of q^+ if $0 < q^+ < 1$. It would be of interest also to find the upper bound for each orbit and to see how the upper bound varies with respect to the right/left states of q^+. Indeed, by (5.12), we get a unique critical point $q^+ = \theta^*$ such that $\frac{dp}{dq^+}|_{q^+=\theta^*} = 0$. The second derivative of p with respect to q^+ is

$$\frac{d^2p}{dq^{+2}} = -\frac{\mu s}{2\kappa}\left[\frac{1}{(c+s)(1-q^+)^2} + \frac{1}{(s-c)q^{+2}}\right], \tag{5.17}$$

Noting that $0 \le c < s$, it is easy to verify that $\frac{d^2p}{dq^{+2}} < 0$ at $q^+ = \theta^*$. Moreover, we know that $p(q_l^+) = p(q_r^+) = 0$. Hence p attains the maximal value at $q^+ = \theta^*$ given by

$$p_{\max} = \frac{\mu s}{2\kappa}\left[\frac{\ln(1-\theta^*)}{c+s} - \frac{\ln\theta^*}{c-s}\right] + \sigma, \tag{5.18}$$

where

$$\sigma = -\frac{\mu s}{2\kappa}\left[\frac{\ln(1-q_l^+)}{c+s} - \frac{\ln q_l^+}{c-s}\right], \qquad \theta^* = \frac{c+s}{2s}. \tag{5.19}$$

Remark 5. From the above equation, we know that the upper bound $p_{\max}$ of p depends on the left states q_l^+ of q. Also, we can easily verify that upper bound $p_{\max}$ increases with respect to $q_l^+ > \theta^*$ (see Figure 4).

Remark 6. The results obtained above for traveling waves are valid only for the case of directed tissue. For undirected tissue, traveling waves with $c < s$ do not exist. Indeed, in the undirected case, we know that $q^+ = q^- = \frac{1}{2}$, and the system (5.6) is reduced to a scalar equation

$$p_z = \frac{\mu^2}{c^2 - s^2}cp. \tag{5.20}$$

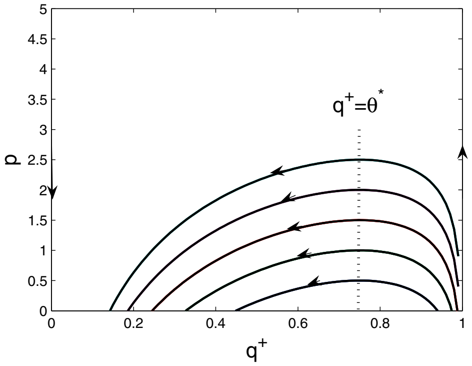


FIG. 4. *The illustration of a family of heteroclinic orbits for the system* (5.6)*, where* $c = 1, s = 2, \mu = 2, \kappa = 1,$ *and* $\theta^* = 0.75$*. The arrow denotes the orientation of trajectories to the system* (5.6)*.*

Clearly, equation (5.20) has no solution satisfying boundary conditions (5.3).

Remark 7. The situation of nested heteroclinic orbits which correspond to traveling waves is also known from other biological applications, for example, for an epidemic with moving infectives (see [24]).

6. Conclusions.

In this study, we analyze the one-dimensional mesenchymal motion model proposed by Hillen [12]. We establish the global existence of classical solutions for both cases of directed and undirected tissue. Particularly, we show that the model of undirected tissue ($\kappa = 0$) has a constant solution for fiber orientation distribution such that $q(t, x, +s) = q(t, x, -s) = \frac{1}{2}$, which means cells have no preference in choosing a particular movement direction and they have equal probability of moving to the right or left. We discuss the existence of inhomogeneous steady states for the case of directed tissue and identify a mechanism of cell aggregation. We rigorously show the convergence of macroscopic limits of the model; i.e., the solution of the mesoscopic model converges to that of the corresponding macroscopic continuum model. Moreover, we study the traveling wave solutions and establish the existence of a traveling pulse in total cell population $p(t, x)$ and traveling front waves in fiber orientation distribution $q^{\pm}(t, x)$. The standing wave ($c = 0$) is admitted in our analysis. This is not surprising considering the fact that cells can move in two directions (left and right) and two traveling waves with opposite direction can eliminate each other to result in a standing wave. All our results are fairly consistent with the biological relevance discussed in paper [12].

The one-dimensional model appears artificial when compared to the real three-dimensional process of cell movement in fiber tissue. The benefit of studying the one-dimensional model in detail is twofold. First of all, this model and its properties give good intuition into mechanisms that might be important in the higher dimensional case. For example, the existence of nonhomogeneous steady states will also be expected for higher dimensional models. Also, the model with directed fibers seems to have a richer behavior. Essentially we identify three distinctions between directed and undirected tissue which are hard to see from the three-dimensional model. We show that for the one-dimensional model, there is no aggregation for undirected tissue, whereas aggregation is possible for directed tissue. In addition, for the macroscopic limit, there are no constraints of convergence for the model of undirected tissue. However, some suitable restriction is needed for directed tissue. Moreover, the model of undirected tissue does not admit traveling waves and the model of directed tissue does. All these distinctions might be true for higher dimensional models.

Second, the model considered here can be used to describe cell movement in highly aligned tissue. In fact, many tissues show a predominant orientation; for example, the rapid spread of glioma cells across the *corpus callosum* results from the migration of individual glioma cells along the highly aligned white matter tracks inside brain tissue [1]. F-actin filaments in vascular smooth muscle cells (VSMCs) are highly aligned on textured polydimethylsiloxane (PDMS) scaffolds [23], and skeletal muscles have a highly organized structure which consists of parallel bundles of multinucleated myotubes that are formed by the fusion of myoblast satellite cells [25]. The model studied here can be used to describe spread and propagation of cells along those aligned tissues. In that case, the traveling pulse waves shown in section 5 correspond to an application of a "comb" to the tissue which is aligned positively or negatively in a common direction. If a brush is applied upstream, say, the fibers will be flipped and higher alignment to the right results, we call these waves *alignment waves*; see also our simulations in Figures 1–3.

For the application of these models to cancer invasion through collagen tissue, the undirected formalism is important. The result of no traveling pulses for that case does not preclude invasions. It precludes only invasion in a self-similar fashion. It is still possible that cells invade new areas, in particular if nonlinear proliferation terms are added. The existence of traveling waves under incorporation of cell proliferation is an interesting open question that comes out of the research done here.

Mathematically, the higher dimensional mesenchymal motion models show significant differences when compared to the one-dimensional case. In one dimension, fiber orientation $q(t, x, \theta)$ has only two directions and hence is bounded due to the normalization condition $q^+ + q^- = 1$. However, in higher dimensions, fibers have infinitely many directions, and highly aligned tissue corresponds to $q(t, x, \theta)$ being a Dirac delta function along that direction. Hence the function spaces have to be chosen to include nonintegrable distributions, and standard L^2 or L^∞ methods do not apply. In a forthcoming paper [13], we will study the existence of solutions for the high dimensional mesenchymal motion models in a Banach space of measurable functions using semigroup theory. If the existence theory stands, we can look into the interesting network formation dynamics, which were found numerically in Painter [21].

Acknowledgments. We thank Dr. D. Wrzosek and Dr. K. P. Hadeler for stimulating discussions and suggestions. We also thank an anonymous referee for remarks that helped to improve the manuscript.

REFERENCES

[1] A. C. Bellail, S. B. Hunter, D. J. Brat, C. Tan, and E. G. Van Meir, *Microregional extracellular matrix heterogeneity in brain modulates glioma cell invasion*, Int. J. Biochem. Cell Biol., 36 (2004), pp. 1046–1069.

[2] A. Bressan, *Hyperbolic System of Conservation Laws: The One-Dimensional Cauchy Problem*, Oxford University Press, New York, 2001.

[3] A. Chauviere, T. Hillen, and L. Preziosi, *Modeling cell movement in anisotropic and heterogeneous network tissues*, Networks and Heterogeneous Media, 2 (2007), pp. 333–357.

[4] A. Chauviere, T. Hillen, and L. Preziosi, *Modeling the motion of a cell population in the extracellular matrix*, Discrete Contin. Dyn. Syst., Dynamical Systems and Differential Equations. Proceedings of the 6th AIMS International Conference, Suppl., (2007), pp. 250–259.

[5] R. Erban and H. Othmer, *From signal transduction to spatial pattern formation in E. coli: A paradigm for multiscale modeling in biology*, Multiscale Model. Simul., 3 (2005), pp. 362–394.

[6] R. Erban and H. Othmer, *Taxis equations for amoeboid cells*, J. Math. Biol., 54 (2007), pp. 847–885.

[7] P. Friedl and E. B. Bröcker, *The biology of cell locomotion within three dimensional extracellular matrix*, Cell. Mol. Life Sci., 57 (2000), pp. 41–64.

[8] S. Goldstein, *On diffusion by discontinuous movements and the telegraph equation*, Quart. J. Mech. Appl. Math., 4 (1951), pp. 129–156.

[9] K. P. Hadeler, *Reaction transport systems in biological modelling*, in Mathematics Inspired by Biology, Lecture Notes in Math. 1714, V. Capasso and O. Diekmann, eds., Springer-Verlag, Heidelberg, 1999, pp. 95–150.

[10] J. P. Hespanha, *Uniform stability of switched linear systems: Extensions of LaSalle's invariant principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–482.

[11] T. Hillen, *Invariant principle for hyperbolic random walk systems*, J. Math. Anal. Appl., 210 (1997), pp. 360–374.

[12] T. Hillen, *M^5 mesoscopic and macroscopic models for mesenchymal motion*, J. Math. Biol., 53 (2006), pp. 585–616.

[13] T. Hillen, P. Hinow, and Z. A. Wang, *Measure-valued solutions for a kinetic model of cell movement in network tissues,* submitted.

[14] T. Hillen, C. Rohde, and F. Lutscher, *Existence of weak solutions for a hyperbolic model for chemosensitive movement*, J. Math. Anal. Appl., 260 (2001), pp. 173–199.

[15] T. Hillen and A. Stevens, *Hyperbolic models for chemotaxis in* 1*-D*, Nonlinear Anal. Real World Appl., 1 (2000), pp. 409–433.

[16] H. J. Hwang, K. Kang, and A. Stevens, *Global existence of classical solutions for a hyperbolic chemotaxis model and its parabolic limit*, Indiana Univ. Math. J., 55 (2006), pp. 289–316.

[17] M. Kac, *A stochastic model related to the telegrapher's equation*, Rocky Mountain J. Math., 4 (1956), pp. 497–509.

[18] J. P. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 25, SIAM, Philadelphia, 1976.

[19] P. G. LeFloch, *Hyperbolic Systems of Conservation Laws: The Theory of Classical and Nonclassical Shock Waves*, Lectures in Math., ETH Zürich, Birkhäuser Verlag, Basel, 2002.

[20] H. Matano, *Convergence of solutions of one-dimensional semilinear parabolic equations*, J. Math. Kyoto Univ., 18 (1978), pp. 221–227.

[21] K. Painter, *Modelling cell migration strategies in the extracellular matrix*, J. Math. Biol., 2008 (electronic).

[22] J. C. Robinson, *Infinite-Dimensional Dynamical Systems: An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*, Cambridge University Press, Cambridge, UK, 2001.

[23] S. Sarkar, D. Manisha, P. Rourke, T. A. Desai, and J. Y. Wong, *Vascular tissue engineering: Microtextured scaffold templates to control organization of vascular smooth muscle cells and extracellular matrix*, Acta Biomater., 2005, pp. 93–100.

[24] N. Shigesada and K. Kawasaki, *Biological Invasions: Theory and Practice*, Oxford University Press, Oxford, UK, 1997.

[25] P. M. Wigmore and G. F. Dunglison, *The generation of fiber diversity during myogenesis*, J. Dev. Biol., 42 (1998), pp. 117–25.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 398–418

DIFFRACTION BY A NONCONVEX POLYGON*

BAIR V. BUDAEV[†] AND DAVID B. BOGY[†]

Abstract. The probabilistic approach to wave propagation and diffraction is applied to a typical problem of diffraction by a nonconvex polygon. The solution is obtained using a transparent technique that employs a floating coordinate system, and it combines ideas from ray theory, stochastic analysis, and complex analysis. The obtained solution is compatible with intuitive ideas about diffraction, and it admits simple implementations.

Key words. diffraction, scattering, ray theory, Feynman–Kac formula, wave propagation

AMS subject classifications. 78A40, 78A45, 74J20

DOI. 10.1137/070699548

1. Introduction.

It is widely known that the process of wave propagation may be conveniently separated into (a) propagation along the rays, which makes possible the long-range transport of energy, and (b) diffusion across the rays, which smooths the distribution of energy and spreads it to the shadow zones. These two processes correspond to different physical phenomena, and they are described by very different mathematical models. Thus, the ray theory [15], describing propagation along the rays, involves first-order partial differential equations, and it provides approximate solutions of the wave equation which have a clear mathematical structure as well as a simple physical meaning. In contrast, the process of diffraction (diffusion across the rays) is described by second-order partial differential equations which are difficult for analysis but, if solved, provide a correction of the ray theory approximation to the exact solution of the considered problem of wave propagation.

The difference in the mathematical foundations of ray and diffraction theories is reflected in the difference in the levels of their development. The ray approximation to the solutions of the wave equation can be computed by a canonical procedure which remains valid in very general settings, and it admits a clear interpretation in physically meaningful terms. However, computation of the diffracted fields remains a difficult problem which has to be studied by special methods tailored to each particular configuration. Moreover, in most cases for which the solutions of the problem of diffraction are known, they either present extremely complicated analytic expressions or rely on massive computations which do not mimic any physical processes and, therefore, do not extend the understanding of the wave propagation phenomena.

For more than one hundred years since the first problems of diffraction were formulated, exact descriptions of the diffracted fields have been obtained only for a small number of domains bounded by surfaces of simple shapes, such as a wedge, cylinder, sphere, or ellipsoid, and most of these solutions have been obtained either by the method of separation of variables or by integral transform methods leading to functional or integral equations solvable by the Wiener–Hopf or other analytic

*Received by the editors August 7, 2007; accepted for publication (in revised form) July 14, 2008; published electronically November 5, 2008. This research was supported by NSF grant CMS-0408381, DOE contract DE-AC02-05CH11231, and by the William S. Floyd, Jr., Distinguished Professorship in Engineering held by D. Bogy.

http://www.siam.org/journals/siap/69-2/69954.html

[†]Department of Mechanical Engineering, Etcheverry Hall, MC 1740, University of California Berkeley, Berkeley, CA 94720 (budaev@berkeley.edu, dbogy@cml.me.berkeley.edu).

methods [4, 12]. However, despite their limited areas of applicability, these exact solutions of particular problems of diffraction by simple objects play an important role in the further understanding of wave propagation, and they have also been used as building blocks for various iterative techniques [1, 2, 3, 14, 16] developed for the analysis of diffraction by more complex objects, such as polygons, for example.

The above-mentioned analytic approaches to the computation of diffracted fields were recently complemented by a probabilistic method which represents solutions of certain second-order partial differential equations by the Feynman–Kac formulas [11, 13], which are mathematical expectations of specific functionals depending on the trajectories of Brownian motions. This method has already been successfully applied to a number of recognized difficult problems, including the problems of diffraction by a finite segment [6], by a half-plane with piecewise constant impedance, and by an arbitrary convex polygon with sidewise constant surface impedance [8]. These problems do not have simple closed-form solutions, but their probabilistic solutions are transparent, simple to implement, and provide clear interpretation of diffraction in terms of diffusion across the rays. The probabilistic approach to wave propagation appears as an extension of the ray method approximation to the exact solutions, which suggests that the probabilistic solutions may be used not only at high frequencies but also at intermediate and low frequencies.

The probabilistic method presented in [8] made it possible to obtain theoretically exact representations for wave fields in the exterior of an arbitrary convex polygon, but due to some technical reasons it could not be directly applied to problems in the exterior of a nonconvex polygon. However, in a more recent development the probabilistic approach has been generalized and enhanced to the extent which makes it possible to use it for the description of wave fields in domains of general shape, including domains bounded by nonconvex polygons. This extension is discussed in [10] in a very general setting which may be excessive for a transparent presentation of its main ideas. For this reason, here we do not directly use the results from [10] but rederive them in a simplified form just sufficient for the analysis of a particular problem of diffraction by a nonconvex polygon. We hope that such an approach will be helpful for the further demonstration of the capabilities of the probabilistic approach to wave propagation.

This paper is organized as follows: two introductory sections, 2 and 3, are followed by the technical sections, 4–6, which lead to the solution of the main problem of diffraction in sections 7 and 8.

In section 2, we introduce notation for handling multiple systems of polar coordinates which are simultaneously used for the adequate description of wave fields in the exterior of a polygon. The lack of a coordinate system naturally associated with the exterior of a polygon seems to be a major obstacle for the description of wave fields in such domains, but the use of a floating coordinate system makes it possible to get around this obstacle by selecting different coordinate systems for different observation points. In section 3 the problem of diffraction by a nonconvex polygon is reduced to the fundamental Problem-1, the computation of wave fields that have specified jumps along rays originating from the vertices of the polygon.

The main technical tools of the paper are developed in section 4, which deals with the problem of radiation into a wedge with virtually arbitrary boundary conditions. The obtained representation generalizes the results from both [8] and [10], which are restricted to analytic boundary conditions and to wedges with angles smaller than 270°, respectively. The results of section 4 are then utilized in sections 5 and 6 as building blocks for the solutions of the radiation problems in the exterior of convex and nonconvex polygons, which are covered by sets of overlapping wedges. Finally, in section 7 we obtain the solution of Problem-W for a nonconvex polygon, and in section 8 the feasibility of the obtained expressions is confirmed by numerical examples.

2. Notation and coordinate systems. Let Γ be a nonconvex N-sided polygon with the vertices $O_0, O_1, \ldots, O_{N-1}$, shown in Figure 1 for the case when $N = 6$. The polygon Γ subdivides the plane into two domains, but our interest here is restricted to the exterior domain G.

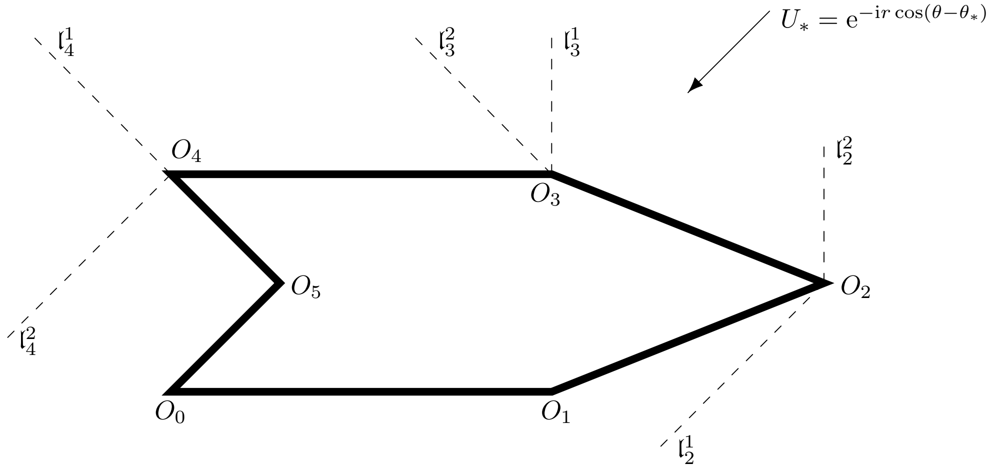


FIG. 1. *Geometry of the problem.*

To describe wave fields in the exterior of a polygon it is convenient to use several coordinate systems simultaneously. The standard polar coordinates (r, θ) are used as a universal reference system, but we also use polar coordinates (r_n, θ_n) centered at the vertices O_n and calibrated by the conditions

$$\theta_n(O_{n-1}) = 0, \tag{2.1}$$

which become meaningful for any integer n if we adopt the periodicity convention

$$O_{n+N} \equiv O_n. \tag{2.2}$$

Similar conventions will also be applied to the exterior angle α_n at the vertex O_n, as well as the sides $\mathfrak{g}_n^\pm$ of the polygon and their lengths $L_n^\pm$ defined as

$$\mathfrak{g}_n^+ = \mathfrak{g}_{n+1}^- = O_n O_{n+1} \qquad \text{and} \qquad L_n^+ = L_{n+1}^- = |O_n O_{n+1}|, \tag{2.3}$$

respectively. It should be noted that although the use of the overlapping notations $\mathfrak{g}_n^+ = \mathfrak{g}_{n+1}^-$ and $L_n^+ = L_{n+1}^-$ may initially look confusing, it is justified by the duality of the position of the segment $O_n O_{n+1}$, which may be considered either as the side attached to the vertex O_n from the right or as the side attached to the vertex O_{n+1} from the left.

To eliminate the ambiguity caused by the use of several coordinate systems it is convenient to include reference to the coordinate system in the notation of the functions. Thus, a function in the domain G is denoted hereafter either as $F(r, \theta)$ or as $F(r, \theta; n)$. In the first case, the pair (r, θ) represents an observation point P in the

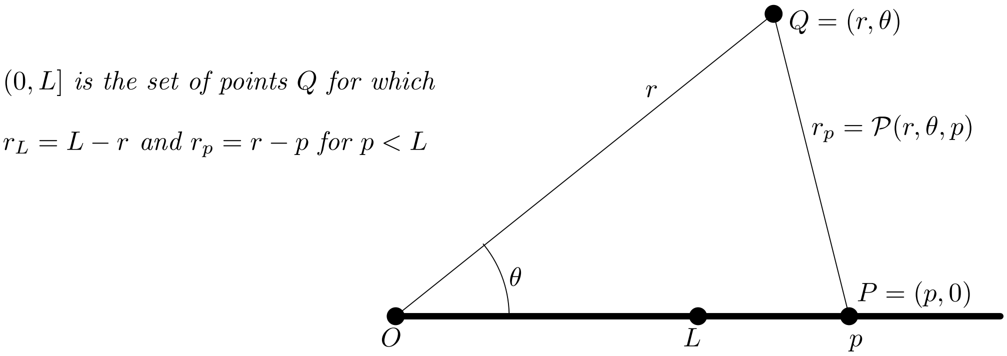

$(0,L]$ is the set of points Q for which
$r_L = L - r$ and $r_p = r - p$ for $p < L$
$Q = (r, \theta)$
r
$r_p = \mathcal{P}(r, \theta, p)$
θ
$P = (p, 0)$
O
L
p

FIG. 2. *Analytic description of the segment* $(0, L]$.

standard reference system. In the second case, the index n indicates that (r, θ) are the coordinates of P in the nth coordinate system.

When different coordinates (r_n, θ_n) refer to the same point of the plane there is a one-to-one correspondence between any pair of such coordinates. For example, applying the cosine theorem to the triangles $O_{n-1}PO_n$ and O_nPO_{n+1}, where P is the observation point, it is easy to show that the coordinates r_{n+1} and r_{n-1} are related to (r_n, θ_n) through the analytic expressions

$$r_{n-1} = \mathcal{P}(r_n, \theta_n, L_n^-), \qquad r_{n+1} = \mathcal{P}(r_n, \alpha_n - \theta_n, L_n^+), \tag{2.4}$$

$$\theta_{n-1} = \alpha_{n-1} - \mathcal{F}(r_n, \theta_n, L_n^-), \qquad \theta_{n+1} = \mathcal{F}(r_n, \alpha_n - \theta_n, L_n^+), \tag{2.5}$$

where

$$\mathcal{P}(r, \theta, L) = \sqrt{r^2 - 2rL\cos\theta + L^2} \equiv \sqrt{(r\mathrm{e}^{\mathrm{i}\theta} - L)(r\mathrm{e}^{-\mathrm{i}\theta} - L)} \tag{2.6}$$

and

$$\mathcal{F}(r, \theta, L) = \arcsin\left(\frac{r\sin\theta}{\mathcal{P}(r, \theta, L)}\right). \tag{2.7}$$

Functions $\mathcal{P}(r, \theta, L)$ and $\mathcal{F}(r, \theta, L)$ are also useful for analytical continuation of "real" geometric objects to the complex space. For example, let $I(\phi; 0, L)$ and $I(\phi; L, \infty)$, where

$$I(\phi, a, b) = \{r, \theta : \ \theta = \phi;\ a < r \leq b\}, \tag{2.8}$$

be a segment and a half-line which comprise the ray $r > 0$, $\theta = \phi$. Then, applying the cosine theorem to the triangle $\triangle OPQ$ shown in Figure 2, we conclude that if $Q \in (0, L]$, then $|QP| = p - r$ for all $p \geq L$, and therefore $I(\phi; 0, L)$ can be described by the formula

$$(r, \theta) \in I(\phi; 0, L) \iff \begin{cases} \mathcal{P}(r, \theta - \phi, p) = p - r & \text{if} \quad p = L, \\ \mathcal{P}(r, \theta - \phi, p) = r - p & \text{if} \quad p < L, \end{cases} \tag{2.9}$$

which provides analytic continuation of the intervals $I(\phi; 0, L)$ to a surface in the complex space formed by the pairs (r, θ).

3. The problem. Let a plane incident wave arriving from infinity in the domain G be defined as

$$U_*(r,\theta) = \mathrm{e}^{-\mathrm{i}kr\cos(\theta-\theta_*)}. \tag{3.1}$$

Then, the problem of diffraction of this wave by the perfectly reflecting polygon Γ can be formulated as the problem of computation of the total field U which is bounded in G and satisfies the Helmholtz equation

$$\nabla^2 U + k^2 U = 0 \qquad \text{in} \quad G, \tag{3.2}$$

complimented by the Dirichlet boundary conditions $U|_\Gamma = 0$, and by the requirement that $U(r,\theta)$ admit the decomposition

$$U = U_g + U_d \tag{3.3}$$

into the predefined piecewise continuous geometric component $U_g(r,\theta)$ and the unknown piecewise continuous scattered field $U_d(r,\theta)$, which has the following asymptote

$$U_d(r_\mu,\theta_\mu;\mu) = \frac{f(\theta)\mathrm{e}^{\mathrm{i}kr_\mu}}{\sqrt{kr_\mu}} + o(1/\sqrt{kr_\mu}), \qquad \theta_\mu \neq \phi_\mu^\nu, \tag{3.4}$$

everywhere in G, except for a finite number of predefined semiaxes

$$\mathfrak{l}_\mu^\nu : \qquad r_\mu \geq 0, \quad \theta_\mu = \phi_\mu^\nu, \tag{3.5}$$

along which the geometric field $U_g(r,\theta)$ has jumps. To keep track of these half-lines we use a double-index notation $\mathfrak{l}_\mu^\nu$, where the lower index corresponds to the vertex O_μ from which the ray originates and the upper index enumerates the rays originating from this vertex.

The geometric field $U_g(r,\theta)$ can be explicitly defined by a straightforward application of the laws of geometrical optics which, however, are easier to apply than to formalize. To keep focused on the principal issues we illustrate the structure of the geometric field for the simple but representative configuration shown in Figure 1. In this case the incident wave illuminates faces $\mathfrak{g}_2^+ = O_2O_3$ and $\mathfrak{g}_3^+ = O_3O_4$ of the polygon, and the waves reflected from these faces do not illuminate any other parts of the polygon, so that the geometric field $U_g(r,\theta)$ admits the decomposition

$$U = U_i + U_r^2 + U_r^3 \tag{3.6}$$

into the incident wave U_i and the waves U_r^2 and U_r^3 reflected by the sides $\mathfrak{g}_2^+$ and $\mathfrak{g}_3^+$, respectively. These waves have the following piecewise continuous structure:

$$U_i(r,\theta) = \begin{cases} \mathrm{e}^{-\mathrm{i}kr\cos(\theta-\theta_*)} & \text{between } \mathfrak{l}_2^1 \text{ and } \mathfrak{l}_4^2 \text{, counterclockwise,} \\ 0 & \text{everywhere else,} \end{cases} \tag{3.7}$$

and

$$U_r^\mu(r,\theta) = \begin{cases} K\mathrm{e}^{-\mathrm{i}kr\cos(\theta+\theta_*^\mu)-\mathrm{i}k\lambda_\mu} & \text{between } \mathfrak{l}_\mu^2 \text{ and } \mathfrak{l}_{\mu+1}^1 \text{, counterclockwise,} \\ 0 & \text{everywhere else,} \end{cases} \tag{3.8}$$

where $\mu = 2, 3$ are the indices of the sides illuminated by the incident wave,

$$K = -1 \tag{3.9}$$

is the reflection coefficient corresponding to the considered Dirichlet boundary conditions, and $\lambda_\mu = r(O_\mu)\cos[\theta(O_\mu) - \theta_*]$ is the phase of the incident wave at the vertex O_μ.

It is easy to see that the geometric field obeys all of the conditions of the problem of diffraction except that it is not continuous along a finite number of the rays $\mathfrak{l}_\mu^\nu$, which can be identified by an elementary geometric optical analysis. Accordingly, the scattered field U_d should admit the decomposition

$$U_d(r,\theta) = \sum_{\mu=1}^{N}\sum_{\nu=1}^{N_\mu} U_\mu^\nu(r,\theta;\phi_\mu^\nu), \tag{3.10}$$

into the sum of $\widetilde{N} = \sum N_\mu$ diffracted fields U_μ^ν each of which compensates the jump of the geometric field along one and only one semiaxis $\mathfrak{l}_\mu^\nu$. Next, assuming that every individual diffracted field is 2π-periodic with respect to the angular coordinate, we arrive at the decomposition

$$U_\mu^\nu(r,\theta;\phi) = K_\mu^\nu \mathrm{e}^{-\mathrm{i}k\lambda_\mu} \sum_{j=-\infty}^{\infty} U_{\mu,\phi}(r,\theta + 2\pi j), \tag{3.11}$$

where $U_{\mu,\phi}(r,\theta)$ is the solution of the following problem.

Problem-1. *Find a bounded solution $U_{\mu,\phi}(r,\theta)$ of the Helmholtz equation $\nabla^2 U_{\mu,\phi} + k^2 U_{\mu,\phi} = 0$ which is defined in the domain $r > 0$, $-\infty < \theta < \infty$, obeys the boundary conditions $U_{\mu,\phi}|_{\mathfrak{g}_m} = 0$ for all integers m, has an asymptote*

$$U_{\mu,\phi}(r_\mu,\theta_\mu;\mu)\mathrm{e}^{-\mathrm{i}kr_\mu} = o(1), \qquad r_\mu \to \infty, \quad \theta_\mu \neq \phi,$$

and satisfies the interface conditions

$$U_{\mu,\phi}(r_\mu,\phi + 0;\mu) - U_{\mu,\phi}(r_\mu,\phi - 0;\mu) = \mathrm{e}^{\mathrm{i}kr_\mu}, \qquad \left.\frac{\partial U_{\mu,\phi}}{\partial \theta_\mu}\right|_{\theta_m = \phi + 0} = \left.\frac{\partial U_{\mu,\phi}}{\partial \theta_\mu}\right|_{\theta_m = \phi - 0},$$

formulated in the μth coordinate system (r_μ,θ_μ).

As shown above, the formulas (3.10) and (3.11) reduce the problem of diffraction to the fundamental Problem-1, which is formulated in the domain $-\infty < \theta < \infty$ and has to be solved with several sets of the parameters μ and ϕ.

4. Radiation into a wedge. Our approach to Problem-1 is based on obtaining a simple representation of the solution of the following basic problem of wave radiation into a wedge.

Problem-*W*. *Find a solution $U(r,\theta)$ of the Helmholtz equation which is bounded in a wedge $r > 0$, $\alpha_1 < \theta < \alpha_2$, has the asymptote $\mathrm{e}^{-\mathrm{i}kr}U(r,\theta) = o(1)$ as $r \to \infty$, and satisfies the boundary conditions*

$$U(r,\alpha_1) = f(r,\alpha_1), \qquad U(r,\alpha_2) = f(r,\alpha_2), \tag{4.1}$$

where

$$f(r,\theta) = \begin{cases} f_1(r) & \text{if} \quad \theta = \alpha_1, \\ f_2(r) & \text{if} \quad \theta = \alpha_2, \end{cases} \tag{4.2}$$

is a "boundary function" defined only on the faces $\theta = \alpha_1$ and $\theta = \alpha_2$.

This problem was studied in detail in [8, 10], but since the results obtained there are not yet commonly known, we briefly reproduce them here in a form adapted to our current needs.

It is shown in [5] that if both $f_1(r)\mathrm{e}^{-\mathrm{i}kr}$ and $f_2(r)\mathrm{e}^{-\mathrm{i}kr}$ are analytic and bounded in the complex r-domain $0 < \arg(r) < \pi/2$, then $U(r, \theta)$ admits the probabilistic representation

$$U(r,\theta) = \mathrm{e}^{\mathrm{i}kr}\mathbf{E}\left\{f(\xi_\tau, \eta_\tau)\mathrm{e}^{\mathrm{i}k[S(\tau)-\xi_\tau]}\right\}, \qquad S(\tau) = \frac{1}{2}\int_0^\tau \xi_t \mathrm{d}t, \tag{4.3}$$

where $\mathbf{E}$ denotes the mathematical expectation computed over the trajectories of the radial and angular motions ξ_t and η_t that are controlled by the stochastic equations

$$\xi_0 = r, \qquad \mathrm{d}\xi_t = \xi_t \mathrm{d}w_t^1 + \xi_t\left(\frac{1}{2} + \mathrm{i}k\xi_t\right)\mathrm{d}t, \tag{4.4}$$

$$\eta_0 = \theta, \qquad \mathrm{d}\eta_t = \mathrm{d}w_t^2, \tag{4.5}$$

and stopped at the exit time $t = \tau$ defined as the first time when the angular motion η_t eventually hits one of the faces $\eta_t = \alpha_1$ or $\eta_t = \alpha_2$. Obviously, the angular motion η_t is contained in the segment $0 \leq \eta_t \leq \alpha$, while the radial motion ξ_t at any $t > 0$ runs inside the first quarter $0 \leq \arg(\xi_t) < \pi/2$ drifting to an unreachable point $\xi = \mathrm{i}/2k$.

Although solution (4.3) is very convenient, it is of limited use because it can only be applied to the cases when the function $f(\xi, \eta)$ is analytic with respect to the first argument ξ. However, using (4.3), it is possible derive a representation of the the field $U(r, \theta)$ in much less restrictive form,

$$U(r,\theta) = \mathrm{e}^{\mathrm{i}kr}\mathbf{E}\left\{f(\hat{\xi}_\tau, \eta_\tau)\mathrm{e}^{\mathrm{i}k[S(\hat{\xi}_\tau)-\hat{\xi}_\tau]}\right\}, \qquad S(\tau) = \frac{1}{2}\int_0^\tau \hat{\xi}_t \mathrm{d}t, \tag{4.6}$$

where the angular motion η_t runs exactly as in (4.3), while the radial motion $\hat{\xi}_t$ for most of the time runs as the radial motion ξ_t from (4.3) but at the exit time $t = \tau$ jumps to a certain point $\hat{\xi}_\tau \in (0, \infty)$ on the positive semiaxis where the boundary function $f(\xi, \eta)$ is defined so that no analytical continuation of $f(\xi, \eta)$ is required. More precisely, on the time interval $0 < t < \tau$ the motion $\hat{\xi}_t$ is controlled by the stochastic differential equation

$$\hat{\xi}_0 = r, \qquad \mathrm{d}\hat{\xi}_t = \hat{\xi}_t \mathrm{d}w_t^1 + \hat{\xi}_t\left(\frac{1}{2} + \mathrm{i}k\hat{\xi}_t\right)\mathrm{d}t \qquad \text{if} \quad 0 \leq t < \tau, \tag{4.7}$$

and then suddenly, at the exit time $t = \tau$, it moves to the final position

$$\hat{\xi}_\tau = \begin{cases} \inf\left\{\mathfrak{z}_+^1(r,\theta) \cup \mathfrak{z}_-^1(r,\theta)\right\} & \text{if} \quad \eta_\tau = \alpha_1, \\ \inf\left\{\mathfrak{z}_+^2(r,\theta) \cup \mathfrak{z}_-^2(r,\theta)\right\} & \text{if} \quad \eta_\tau = \alpha_2, \end{cases} \tag{4.8}$$

where $\mathfrak{z}_\pm^1(r, \theta)$ and $\mathfrak{z}_\pm^2(r, \theta)$ are the intersections of the trajectories of the auxiliary motions

$$\zeta_\pm^1(t) = \hat{\xi}_t \mathrm{e}^{\pm\mathrm{i}(\eta_t - \alpha_1)} \qquad \text{and} \qquad \zeta_\pm^2(t) = \hat{\xi}_t \mathrm{e}^{\pm\mathrm{i}(\alpha_2 - \eta_t)} \tag{4.9}$$

with the semiaxis $\zeta > 0$. It is clear that the final position $\hat{\xi}_\tau$ of the motion $\hat{\xi}_t$ is determined by the entire trajectory of both components of the two-dimensional motion

$(\hat{\xi}_t, \eta_t)$ launched at $t = 0$ from the point (r, θ) and stopped as soon as the angular motion η_t reaches at the exit time $t = \tau$ either of the endpoints $\eta_\tau = \alpha_1$ or $\eta_\tau = \alpha_2$.

To make the derivation of (4.6) more transparent we limit ourselves to the case when $\alpha_1 = 0$, $\alpha_2 \equiv \alpha \leq 2\pi$, and consider the problem with the special piecewise constant boundary function

$$f(r,\theta) = \begin{cases} a_0 & \text{if} \quad (r,\theta) \in I(0,0,L), \\ a_1 & \text{if} \quad (r,\theta) \in I(0;L,\infty), \\ a_2 \equiv 0 & \text{if} \quad (r,\theta) \in I(\alpha;0,\infty), \end{cases} \tag{4.10}$$

where a_ν and $L > 0$ are given constants. Then, the basic formula (4.3) leads to the representation

$$U(r,\theta) = \mathrm{e}^{\mathrm{i}kr}\mathbf{E}\left\{a_{\nu(\tau)} \exp\left(\frac{1}{2}\int_0^\tau \mathrm{i}k\xi_t \mathrm{d}t - \mathrm{i}k\xi_\tau\right)\right\}, \tag{4.11}$$

where the index

$$\nu(\tau) = \begin{cases} 0 & \text{if} \quad \eta_\tau = 0, \quad \mathcal{P}(\xi_\tau, \eta_\tau, p) = p - \xi_\tau \quad \forall p \geq L, \\ 1 & \text{if} \quad \eta_\tau = 0, \quad \mathcal{P}(\xi_\tau, \eta_\tau, p) \neq p - \xi_\tau \quad \forall p < L, \\ 2 & \text{if} \quad \eta_\tau = \alpha \end{cases} \tag{4.12}$$

depends on the entire trajectory of the motion (ξ_t, η_t) rather than on its final position (ξ_τ, η_τ).

To compute $\nu(\tau)$ by the formula (4.12) it is necessary to trace the value of the radical

$$\mathcal{P}(\xi_t, \eta_t, p) = \sqrt{(\xi_t \mathrm{e}^{\mathrm{i}\eta_t} - p)(\xi_t \mathrm{e}^{-\mathrm{i}\eta_t} - p)} \tag{4.13}$$

along the trajectory of the motion $P_t = (\xi_t, \eta_t)$. To proceed we represent this function as a product

$$\mathcal{P}(\xi_t, \eta_t, p) = \Xi_p(\zeta_+(t))\, \Xi_p(\zeta_-(t)), \qquad \zeta_\pm(t) = \xi_t \mathrm{e}^{\pm \mathrm{i}\eta_t}, \tag{4.14}$$

where $\Xi_p(\zeta) = \sqrt{\zeta - p}$ is the radical with the branch fixed by a slit along the half-line $\zeta > p$.

Let $\mathbb{C}_p^+$ and $\mathbb{C}_p^-$ be the two sheets of the Riemann surface of the radical $\Xi_p(\zeta)$, so that

$$\Xi_p(\zeta) = \pm\sqrt{|\zeta - p|}\mathrm{e}^{\mathrm{i}\,\mathrm{Arg}(\zeta - p)/2}, \qquad \zeta \in \mathbb{C}_p^\pm. \tag{4.15}$$

Since $\mathcal{P}(r, \theta, p)$ is the distance between the observation point (r, θ) and the point $(p, 0)$ on the main axis, we need to select the branches of the radical in such a way as to have $\mathcal{P}(r, \theta, p) \geq 0$. To guarantee this inequality the initial points $\zeta_\pm(0) = r\mathrm{e}^{\pm \mathrm{i}\theta}$ must be located on different sheets $\mathbb{C}_p^\pm$. For definiteness we assume that

$$\zeta_+(0) = r\mathrm{e}^{\mathrm{i}\theta} \in \mathbb{C}_p^+, \qquad \zeta_-(0) = r\mathrm{e}^{-\mathrm{i}\theta} \in \mathbb{C}_p^-, \tag{4.16}$$

which implies the identities

$$\Xi_p(\zeta_\pm(0)) = \sqrt{|r\mathrm{e}^{\pm\mathrm{i}\theta} - p|}\, \exp\left[\frac{\mathrm{i}}{2}\,\mathrm{Arg}(r\mathrm{e}^{\pm\mathrm{i}\theta} - p)\right], \tag{4.17}$$

leading to the required inequality $\mathcal{P}(r, \theta, p) \geq 0$.

To compute $\Xi(\zeta_+(\tau))$ we need to trace the trajectory of the point $\zeta_+(t) = \xi_t \mathrm{e}^{\mathrm{i}\eta_t}$, which starts from $\zeta_+(0) = r\mathrm{e}^{\mathrm{i}\theta} \in \mathbb{C}_p^+$ and stops at the point $\zeta_\tau^+ = \xi_\tau$ in the first quarter $0 < \arg(\zeta_\tau^-) < \pi/2$. The restraints $0 \le \arg(\xi_t) < \pi/2$ and $\eta_t > 0$ imply that $\zeta_-(t)$ never crosses the ray $\arg(\zeta) = 0$ but may cross any of the rays $\arg(\zeta) = 2\pi n$ with $n > 1$ an even number of times, so that the total number of intersections of the trajectory of ζ_t with the half-line $\zeta > 0$ is even, as illustrated in Figure 3.

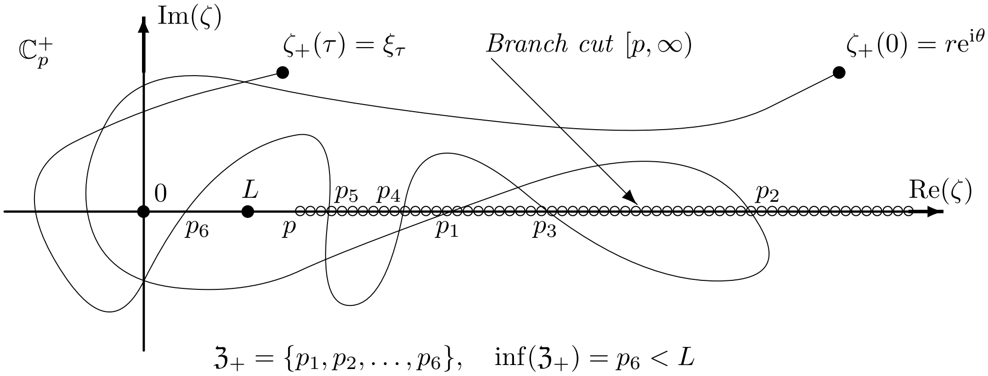

$\mathrm{Im}(\zeta)$
$\mathbb{C}_p^+$
$\zeta_+(\tau) = \xi_\tau$
Branch cut $[p, \infty)$
$\zeta_+(0) = r\mathrm{e}^{\mathrm{i}\theta}$
0
L
p_5
p_4
p_2
$\mathrm{Re}(\zeta)$
p_6
p
p_1
p_3
$\mathfrak{Z}_+ = \{p_1, p_2, \ldots, p_6\}, \quad \inf(\mathfrak{Z}_+) = p_6 < L$

FIG. 3. *Trajectory of* $\zeta_+(t) = \xi_t \mathrm{e}^{\mathrm{i}\eta_t}$ *on the sheet* $\mathbb{C}_p^+$.

Let $\mathfrak{Z}_+$ be the set of all points where the trajectory of $\zeta_+(t)$ intersects the ray $\zeta > 0$. Then, adopting the notation $N(\mathfrak{Z} > p)$ for the number of points of the set $\mathfrak{Z}$ located to the right of p, we conclude that the radical $\Xi_p(\zeta_+(\tau))$ has the value

$$\Xi_p(\zeta_+(\tau)) = \begin{cases} \sqrt{|\xi_\tau - p|}\, \mathrm{e}^{\mathrm{i}\,\mathrm{Arg}(\xi_\tau - p)/2} & \text{if } N(\mathfrak{Z}_+ > p) \text{ is even,} \\ -\sqrt{|\xi_\tau - p|}\, \mathrm{e}^{\mathrm{i}\,\mathrm{Arg}(\xi_\tau - p)/2} & \text{if } N(\mathfrak{Z}_+ > p) \text{ is odd,} \end{cases} \tag{4.18}$$

which is completely determined by the disposition of the parameter p with respect to the set $\mathfrak{Z}_+$ and does not depend on other details of the trajectory of $\zeta_+(t)$. It is interesting to note that if $\alpha < 3\pi/2$, then the set $\mathfrak{Z}_+$ is empty, and therefore the second option in (4.18) never occurs.

To compute $\Xi_p(\zeta_-(t))$ we need to trace the trajectory of the point $\zeta_-(t) = \xi_t \mathrm{e}^{-\mathrm{i}\eta_t}$, which starts from $\zeta_-(0) = r\mathrm{e}^{-\mathrm{i}\theta} \in \mathbb{C}_p^-$ and stops at the point $\zeta_\tau^- = \xi_\tau$ in the first quarter $0 < \arg(\zeta) < \pi/2$. The restraints $0 \le \arg(\xi_t) < \pi/2$ and $\eta_t > 0$ imply that $\zeta_-(t)$ crosses the half-line $\zeta > 0$ an odd number of times, as illustrated in Figure 4. Therefore, assuming that $\mathfrak{Z}_-$ is the set of all points where $\zeta_-(t)$ intersects this ray $\zeta > 0$, we conclude that the value of $\Xi_p(\zeta_-(\tau))$ is determined by the formula

$$\Xi_p(\zeta_-(\tau)) = \begin{cases} -\sqrt{|\xi_\tau - p|}\, \mathrm{e}^{\mathrm{i}\,\mathrm{Arg}(\xi_\tau - p)/2} & \text{if } N(\mathfrak{Z}_- > p) \text{ is even,} \\ \sqrt{|\xi_\tau - p|}\, \mathrm{e}^{\mathrm{i}\,\mathrm{Arg}(\xi_\tau - p)/2} & \text{if } N(\mathfrak{Z}_- > p) \text{ is odd,} \end{cases} \tag{4.19}$$

which is similar to (4.18).

Finally, combining (4.13) with (4.18) and (4.19), we see that

$$\mathcal{P}(\xi_\tau, \eta_\tau, p) = \begin{cases} \xi_\tau - p & \text{if } \eta_\tau = 0, \quad N(\mathfrak{Z}_- \cup \mathfrak{Z}_+ > p) \text{ is even,} \\ p - \xi_\tau & \text{if } \eta_\tau = 0, \quad N(\mathfrak{Z}_- \cup \mathfrak{Z}_+ > p) \text{ is odd,} \end{cases} \tag{4.20}$$

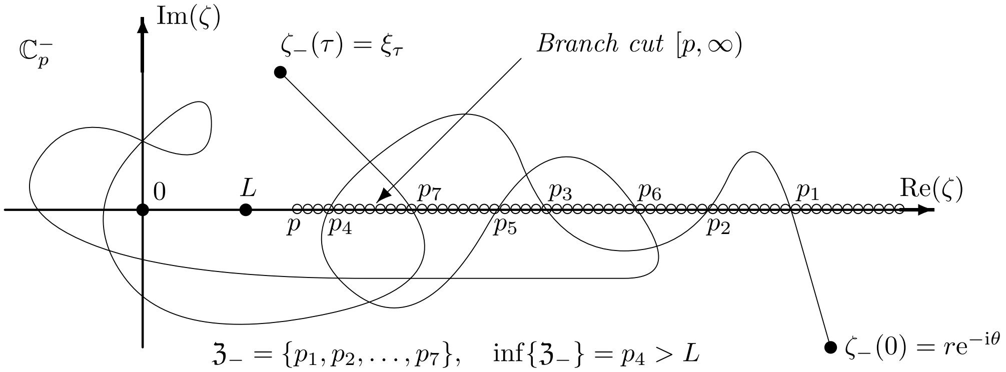

$\mathbb{C}_p^-$
$\mathrm{Im}(\zeta)$
$\zeta_-(\tau) = \xi_\tau$
Branch cut $[p, \infty)$
0
L
p_7
p_3
p_6
p_1
$\mathrm{Re}(\zeta)$
p
p_4
p_5
p_2
$\zeta_-(0) = r\mathrm{e}^{-\mathrm{i}\theta}$
$\mathfrak{Z}_- = \{p_1, p_2, \dots, p_7\}, \quad \inf\{\mathfrak{Z}_-\} = p_4 > L$

FIG. 4. *Trajectory of* $\zeta_-(t) = \xi_t \mathrm{e}^{-\mathrm{i}\eta_t}$ *on the sheet* $\mathbb{C}_p^-$.

and, combining (4.20) with (2.9), we arrive at the remarkable formula

$$\nu(\tau) = \begin{cases} 0 & \text{if} \quad \eta_\tau = 0, \quad \inf\{\mathfrak{Z}_+ \cup \mathfrak{Z}_-\} \geq L, \\ 1 & \text{if} \quad \eta_\tau = 0, \quad \inf\{\mathfrak{Z}_+ \cup \mathfrak{Z}_-\} < L, \\ 2 & \text{if} \quad \eta_\tau = \alpha, \end{cases} \tag{4.21}$$

which makes it possible to evaluate (4.11) by tracing the intersections of the trajectories of $\zeta_\pm(t) = \xi_t \mathrm{e}^{\pm \mathrm{i}\eta_t}$ with the positive semiaxis, but without tracing the radical (4.13).

Formulas (4.3) and (4.11) represent the field $U(r, \theta)$ in the special cases when the boundary function $f(r, \theta)$ is either analytic in the first quarter or a piecewise constant on $\theta = 0$ with only one jump. To obtain the representation of $U(r, \theta)$ corresponding to an arbitrary boundary function $f(r, \theta)$, we first assume that $f(r, \theta)$ has a piecewise constant structure

$$f(r, \theta) = \begin{cases} a_n & \text{if} \quad \theta = 0, \quad r \in I_n = (L_n, L_{n+1}], \qquad n \geq 0, \\ 0 & \text{if} \quad \theta = \alpha, \end{cases} \tag{4.22}$$

where $\{L_n\}$ is an increasing sequence with the first element $L_0 = 0$ and a_n are some constants.

To employ the technique from the previous subsection we observe that the interval I_n admits the representation $I_n = I(0; 0, L_{n+1}) \backslash I(0; 0, L_n)$, where $I(\phi; a, b)$ is the domain in the space (r, θ) defined by (2.9). Then Problem-W can be considered in the domain bounded by the junction $\bigcup I_n$, and its solution (4.6) takes the form

$$U(r, \theta) = \mathrm{e}^{\mathrm{i}kr} \mathbf{E} \left\{ f(L_{\nu(\tau)+1}, \eta_\tau) \exp\left(\frac{1}{2} \int_0^\tau \mathrm{i}k\xi_t \mathrm{d}t - \mathrm{i}k\xi_\tau \right) \right\}, \tag{4.23}$$

where the mathematical expectation is computed over the trajectories of the stochastic processes ξ_t and η_t, which are described by (4.4), (4.5) and are stopped at the exit time τ defined as the first time when $\eta_t = 0$ or $\eta_t = \alpha$. It is easy to see that if $\eta_\tau = 0$, then the exit point $P_\tau = (\xi_\tau, \eta_\tau)$ belongs to the interval $I_{\nu(\tau)}$ with the index

$$\nu(\tau) = n \quad \text{if} \quad L_n < \inf\{\mathfrak{Z}_+ \cup \mathfrak{Z}_-\} \leq L_{n+1}, \tag{4.24}$$

where $\mathfrak{Z}_\pm$ are the intersections of the trajectories of $\zeta_t^\pm = \xi_t \mathrm{e}^{\mathrm{i}\eta_t}$ with the ray $\zeta > 0$. Comparison of the last formula with (4.22) shows that if $\hat{\xi}_\tau = \inf\{\mathfrak{Z}_+ \cup \mathfrak{Z}_-\}$ belongs to the interval I_n, then $\nu_\tau = n$ and $f(\hat{\xi}_\tau, 0) = a_{\nu(\tau)}$. As a result, the solution (4.23) can be converted to the form (4.6), which does not rely on the piecewise structure of $f(r, 0)$, and which, therefore, can be straightforwardly extended to the case when $f(r, 0)$ has a virtually arbitrary structure, particularly to the cases when it is not analytic at all or when it is analytic but grows excessively fast in the first quarter of the complex plane, where the radial motion ξ_t runs.

It is worth noticing that although formulas (4.3) and (4.6) look similar, there is a significant difference between them. Thus, in (4.3) the exit position ξ_τ of the radial motion is always complex, while in (4.6) the exit value $\hat{\xi}_\tau$ is real, which is very important for our purposes.

In the above we considered only the case when the boundary values vanish on the face $\theta = \alpha$, but the case with nonzero boundary values on $\theta = \alpha$ can be considered similarly.

5. Radiation to the exterior of a convex polygon.

The solution of the boundary value problem for the Helmholtz equation in a wedge makes it possible to solve a similar problem of radiation to the exterior of a convex polygon.

Problem-C. *Find the solution $U(r, \theta)$ of the Helmholtz equation $\nabla^2 U + k^2 U = 0$ which is bounded in the exterior of the convex polygon Γ with sides $\mathfrak{g}_n^+$, has the asymptote*

$$U(r, \theta) = u(\theta)\frac{\mathrm{e}^{\mathrm{i}kr}}{\sqrt{kr}} + o\left(\frac{1}{\sqrt{kr}}\right), \qquad r \to \infty,$$

and obeys the boundary condition $U|_{\partial\Gamma} = f$, where $f(r, \theta) = f(r_n, \theta_n; n)$ is a function defined only on the sides $\mathfrak{g}_n^+$ of Γ.

To compute the solution of this problem at an observation point (r, θ) we select an exterior wedge $\mathfrak{O}_\mu$ which contains (r, θ). The boundary of this wedge consists of two sides $\mathfrak{g}_{\mu-1}^+ \equiv \mathfrak{g}_\mu^-$ and $\mathfrak{g}_{\mu+1}^- \equiv \mathfrak{g}_\mu^+$ of the polygon Γ, which are described by the equations

$$\mathfrak{g}_\mu^- = \{\theta_\mu = 0, \quad r_\mu < L_\mu^-\}, \qquad \mathfrak{g}_\mu^+ = \{\theta_\mu = \alpha_\mu, \quad r_\mu < L_\mu^+\}, \tag{5.1}$$

and of two half-lines $\mathfrak{a}_\mu^-$ and $\mathfrak{b}_\mu^+$, which are described by the equations

$$\mathfrak{a}_\mu^- = \{\theta_\mu = 0, \quad r_\mu > L_\mu^-\}, \qquad \mathfrak{a}_\mu^+ = \{\theta_\mu = \alpha, \quad r_\mu > L_\mu^+\}, \tag{5.2}$$

and are shown in Figure 5 by dashed lines.

On the sides described in (5.1) the values of the field $U(r_\mu, \theta_\mu; \mu)$ are preassigned by the boundary conditions, but on the half-lines (5.2) this field may not be known until the problem is solved. However, we assume that the values of $U(r_\mu, \theta_\mu; \mu)$ on the half-lines $\mathfrak{a}_\mu^-$ and $\mathfrak{a}_\mu^+$ are already known, and, using this information, we calculate $U(r_\mu, \theta_\mu; \mu)$ inside $\mathfrak{O}_\mu$ by the formula (4.6), which can be rearranged to the form

$$\begin{aligned} U(r_\mu, \theta_\mu; \mu) = \mathrm{e}^{\mathrm{i}kr_\mu}\mathbf{E}\Big\{&\chi(\tau, t_1) f(\hat{\xi}_\tau, \eta_\tau; \mu)\mathrm{e}^{\mathrm{i}k[S(\tau)-\hat{\xi}_\tau]} \\ &+ \chi_*(\tau, t_1) U(\hat{\xi}_{t_1}, \eta_{t_1}; \mu)\mathrm{e}^{\mathrm{i}k[S(t_1)-\hat{\xi}_{t_1}]}\Big\}, \qquad S(t) = \frac{1}{2}\int_0^t \hat{\xi}_t \mathrm{d}t, \end{aligned} \tag{5.3}$$

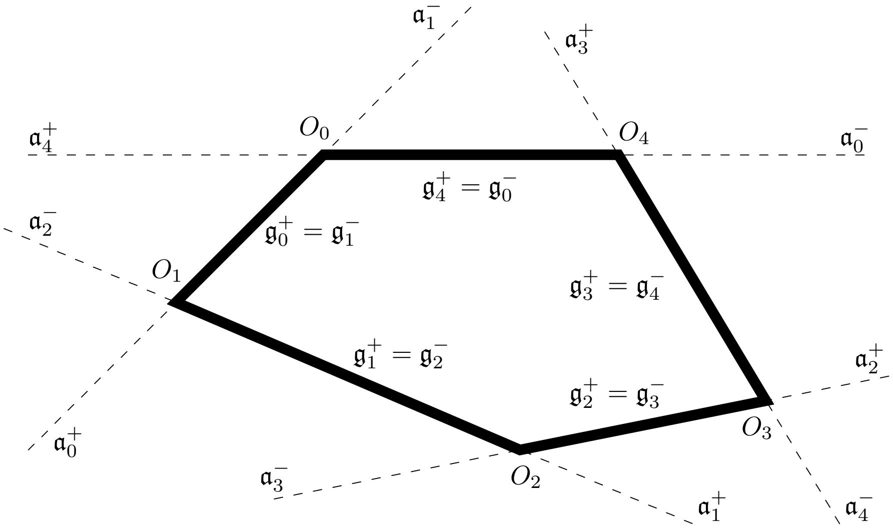


FIG. 5. *Geometry of a convex polygon.*

where

$$\chi(a,b) = \begin{cases} 1 & \text{if} \quad a \geq b, \\ 0 & \text{if} \quad a < b, \end{cases} \qquad \chi_*(a,b) = \begin{cases} 0 & \text{if} \quad a \geq b, \\ 1 & \text{if} \quad a < b, \end{cases} \tag{5.4}$$

and $P_t = (\hat{\xi}_t, \eta_t)$ is a random motion which starts from the point (r_μ, θ_μ) and runs inside $\mathfrak{O}_\mu$ controlled by (4.5) and (4.7)–(4.9) until the earlier of the times $t = \tau$ or $t = t_1$, when it hits either the "real" boundary $\mathfrak{g}_\mu^- \cup \mathfrak{g}_\mu^+$ or the "auxiliary" boundary $\mathfrak{a}_\mu^- \cup \mathfrak{a}_\mu^+$, respectively.

Expression (5.3) cannot be accepted as a solution of the problem because its right-hand side involves yet unknown values $U(\hat{\xi}_{t_1}, \eta_{t_1}; \mu)$ of the field $U(r,\theta)$ on the half-lines $\mathfrak{a}_\mu^\pm$. However, these values can be computed by analogues of the formula (5.3) applied to the neighboring wedges $\mathfrak{O}_{\mu-1}$ and $\mathfrak{O}_{\mu+1}$, the first of which contains $\mathfrak{a}_\mu^-$ while the latter contains $\mathfrak{a}_\mu^+$. To perform such evaluations we observe that the half-lines $\mathfrak{a}_\mu^-$ and $\mathfrak{a}_\mu^+$ are characterized by the equations

$$\mathfrak{a}_\mu^- : \quad r_\mu = r_{\mu-1} + L_\mu^-, \qquad \theta_\mu = \theta_{\mu-1} - \pi + \alpha_\mu^-, \tag{5.5}$$

$$\mathfrak{a}_\mu^+ : \quad r_\mu = r_{\mu+1} + L_\mu^+, \qquad \theta_\mu = \theta_{\mu+1} + \pi - \alpha_\mu^+, \tag{5.6}$$

which show that the ray $\mathfrak{a}_\mu^-$ is a good place for switching from the coordinates (r_μ, θ_μ) to the coordinates $(r_{\mu-1}, \theta_{\mu-1})$, and $\mathfrak{a}_\mu^+$ is good place for switching from (r_μ, θ_μ) to $(r_{\mu+1}, \theta_{\mu+1})$. Then, applying (5.5) and (5.6), we obtain the identities

$$U(\hat{\xi}_{t_1}, \eta_{t_1}; \mu) = U(\hat{\xi}_{t_1} - L_\mu^\pm,\, \eta_{t_1} \pm [\pi - \alpha_\mu^\pm];\, \mu \pm 1), \tag{5.7}$$

with right-hand sides that can be computed by the formula (5.3) adapted to the wedges $\mathfrak{O}_{\mu-1}$ and $\mathfrak{O}_{\mu+1}$. For example, the right-hand part of (5.7), determined by the choice "−," can be represented as the mathematical expectation

$$(5.8)\quad \begin{aligned} &U(\hat{\xi}_{t_1}-L_\mu^-,\, \eta_{t_1}+\alpha_\mu^- - \pi;\, \mu-1) \\ &\quad = \mathrm{e}^{\mathrm{i}k[\hat{\xi}_{t_1}-L_\mu^-]}\mathbf{E}\Big\{\chi(\tau,t_2)f(\hat{\xi}_\tau,\eta_\tau;\mu-1)\mathrm{e}^{\mathrm{i}k[S(\hat{\xi}_\tau)-\hat{\xi}_\tau]} \\ &\qquad\qquad + \chi_*(\tau,t_2)U(\hat{\xi}_{t_2},\eta_{t_2};\mu-1)\mathrm{e}^{\mathrm{i}k[S(\hat{\xi}_{t_2})-\hat{\xi}_{t_2}]}\Big\} \end{aligned}$$

computed over the trajectories of the stochastic process $P_t=(\hat{\xi}_t,\eta_t)$, which is launched at the time $t=t_1+0$ from the initial position

$$(5.9)\qquad \hat{\xi}_{t_1+0}=\hat{\xi}_{t_1}-L_\mu^-,\qquad \eta_{t_1+0}=\eta_{t_1}+\alpha_\mu^- - \pi$$

and runs in the wedge $\mathfrak{O}_{\mu-1}$ until the earlier of the exit times $t=\tau$ or $t=t_2$, when it hits one of the sides $\mathfrak{g}_{\mu-1}^\pm$ of the polygon or their continuations $\mathfrak{a}_{\mu-1}^\pm$.

It is obvious that the value of $U(\xi_{t_2},\eta_{t_2};\mu-1)$ that appears in the right-hand side of (5.8) can be evaluated by a formula similar to that in (5.8). Then, the recursion can be repeated infinitely many times, which eventually results in the expression

$$(5.10)\quad U(r,\theta;\mu)=\mathrm{e}^{\mathrm{i}kr_\mu}\mathbf{E}\Big\{f(\hat{\xi}_\tau,\eta_\tau;n_\tau)\mathrm{e}^{\mathrm{i}k[S(\tau)-\hat{\xi}_\tau]}\Big\},\qquad S(\tau)=\frac{1}{2}\int_0^\tau \hat{\xi}_t\mathrm{d}t-\Lambda_\tau,$$

where the mathematical expectation is computed over trajectories of the stochastic processes n_t, $P_t=(\hat{\xi}_t,\eta_t)$, and Λ_t, which evolve as described below.

The process n_t indicates the index of the currently used coordinate system. It takes integer values which change only at the times $t=t_1,t_2,\ldots$, when P_t reaches one of the rays $\mathfrak{a}_n^\pm$. More precisely, the evolution of n_t is described by the rules

$$(5.11)\qquad n_0=\mu,\qquad n_{t+\mathrm{d}t}=\begin{cases} n_t\pm 1 & \text{if}\quad P_t\in\mathfrak{a}_{n_t}^\pm, \\ n_t & \text{otherwise,}\end{cases}$$

where the initial index μ is not rigidly fixed but must be selected according to the condition that the observation point (r,θ) is located inside the wedge $\mathfrak{O}_\mu$.

The process n_t is closely related to another piecewise process Λ_t described by the equations

$$(5.12)\qquad \Lambda_0=0,\qquad \mathrm{d}\Lambda_t\equiv\Lambda_{t+\mathrm{d}t}-\Lambda_t=\begin{cases} L_{n_t}^\pm & \text{if}\quad P_t\in\mathfrak{a}_{n_t}^\pm, \\ 0 & \text{otherwise,}\end{cases}$$

which show that every time $t=t_\nu$ when the coordinate system is changed, Λ_t takes an increment equal to the distance between the centers of the old and new coordinate systems.

The changes of the coordinate systems at the times $t=t_1$, $t=t_2,\ldots$, affect both of the components of the motion $P_t=(\hat{\xi}_t,\eta_t)$, where η_t is controlled by the stochastic equations

$$(5.13)\qquad \eta_0=\theta,\qquad \eta_{t+\mathrm{d}t}=\eta_t+\mathrm{d}w_t^2+\begin{cases} \alpha_{n_t-1}-\pi & \text{if}\quad P_t\in\mathfrak{a}_{n_t}^-, \\ \pi-\alpha_{n_t} & \text{if}\quad P_t\in\mathfrak{a}_{n_t}^+, \\ 0 & \text{otherwise,}\end{cases}$$

which describe a standard Brownian motion w_t^2 modified by certain jumps at the times $t=t_\nu$. The radial motion $\hat{\xi}_t$ has a more complicated structure. In the intervals

$t \in [t_\nu, t_{\nu+1})$ between changes of the coordinate systems $\hat{\xi}_t$ is a continuous motion governed by the equations

$$\hat{\xi}_0 = r, \qquad \hat{\xi}_{t+\mathrm{d}t} = \hat{\xi}_t + \hat{\xi}_t \mathrm{d}w_t^1 + \hat{\xi}_t \left(\frac{1}{2} + \mathrm{i}k\hat{\xi}_t\right) \mathrm{d}t, \qquad t \leq t_\nu < t_{\nu+1}, \tag{5.14}$$

where w_t^1 is the standard Brownian motion. Then, at the time $t = t_{\nu+1}$ the point $\hat{\xi}_t$ jumps to the position

$$\hat{\xi}_{t_{\nu+1}} = -\mathrm{d}\Lambda_{t_{\nu+1}} + \begin{cases} \inf\left[\mathfrak{Z}^1_+(\hat{\xi}_{t_\nu}, \eta_{t_\nu}) \cup \mathfrak{Z}^1_-(\hat{\xi}_{t_\nu}, \eta_{t_\nu})\right] & \text{if} \quad \eta_t = 0, \\ \inf\left[\mathfrak{Z}^2_+(\hat{\xi}_{t_\nu}, \eta_{t_\nu}) \cup \mathfrak{Z}^2_-(\hat{\xi}_{t_\nu}, \eta_{t_\nu})\right] & \text{if} \quad \eta_t = \alpha_{n_{t-0}}, \end{cases} \tag{5.15}$$

where $\mathrm{d}\Lambda_{t_{\nu+1}}$ is defined by (5.12), while $\mathfrak{Z}^1(\hat{\xi}_{t_\nu}, \eta_{t_\nu})$ and $\mathfrak{Z}^2(\hat{\xi}_{t_\nu}, \eta_{t_\nu})$ are the sets of intersections of the trajectories of the motions

$$\zeta^1_\pm(t) = \hat{\xi}_t \exp[-\mathrm{i}\eta_t] \quad \text{and} \quad \zeta^2_\pm(t) = \hat{\xi}_t \exp[-\mathrm{i}(\alpha_{n_t} - \eta_t)] \qquad t_\nu < t < t_{\nu+1}, \tag{5.16}$$

with the semiaxis $\mathrm{Im}(\zeta) = 0$, $\mathrm{Re}(\zeta) > 0$.

It is important to note that the obtained solution remains valid in the case of an "infinite-sided" polygon Γ with the vertices O_n defined by the recursive process which starts from two initial vertices O_0, O_1 and continues in both directions $n > 1$ and $n < 0$ making steps

$$O_{n+1} = \{r_n = L_n^+,\ \theta_n = \alpha_n\}, \qquad O_{n-1} = \{r_n = L_{n-1}^+,\ \theta_n = \alpha_{n-1}\}, \tag{5.17}$$

determined by the sequences $\{\alpha_\nu\}$ and $\{L_\nu^+\}$, where $L_\nu > 0$, $\alpha_\nu > \pi$, and $\sum(\alpha_\nu - \pi) = \infty$. Assume first that there exists $N > 1$ for which $\sum_{\nu=0}^{N-1}(\alpha_\nu - \pi) = 2\pi$ and the boundary function $f(r, \theta; \nu)$ satisfies the periodicity condition $f(r, \theta; \nu) = f(r, \theta; \nu + N)$. In this case the vertices O_ν with $0 \leq \nu < N - 1$ form an N-sided convex polygon, and formulas (5.10)–(5.16) determine the field radiated into the exterior of this polygon. In the other case, when either of the mentioned conditions is not met, these formulas determine a field radiated to the spiral-like surface bounded by Γ.

6. Radiation to the exterior of a nonconvex polygon. In section 5 the problem of radiation to the exterior of a convex polygon was reduced to the problem of radiation into a wedge. Here a similar idea is employed to obtain the solution of the problem of radiation into the exterior of a nonconvex polygon by recursive applications of already known solutions of the problems of radiation from a convex polygon and from a wedge.

Let Γ be a "single-cavity" nonconvex polygon with the vertices $O_0, \ldots, O_{N-2}$, O_{N-1}, located in such a way that the first $\widetilde{N} \equiv (N-1)$ vertices form a convex polygon $\widetilde{\Gamma}$. This configuration is illustrated in Figure 6, which shows a nonconvex polygon Γ obtained by the addition of a vertex O_5 to the convex pentagon from Figure 5. To keep the notation consistent with the previous sections we reserve the symbols $\mathfrak{g}_n^\pm$ and $\mathfrak{a}_n^\pm$ for the sides $O_n O_{n\pm 1}$ of the polygon Γ and for their continuations beyond the vertices $O_{n\pm 1}$. Similarly, symbols $\widetilde{\mathfrak{g}}_n^\pm$ and $\widetilde{\mathfrak{a}}_n^\pm$ denote the sides and their continuations of the convex polygon $\widetilde{\Gamma}$. This notation is used in Figure 6, which, however, does not show $\widetilde{\Gamma}$-related symbols whenever a similar Γ-related notation can be used.

Looking back at the previous section, we see that the employed technique is built around the ability to represent the exterior G of the convex polygon Γ as a junction $G = \bigcup \mathfrak{O}_n$ of overlapping wedges $\mathfrak{O}_n$ defined as the exterior wedges of the polygon Γ.

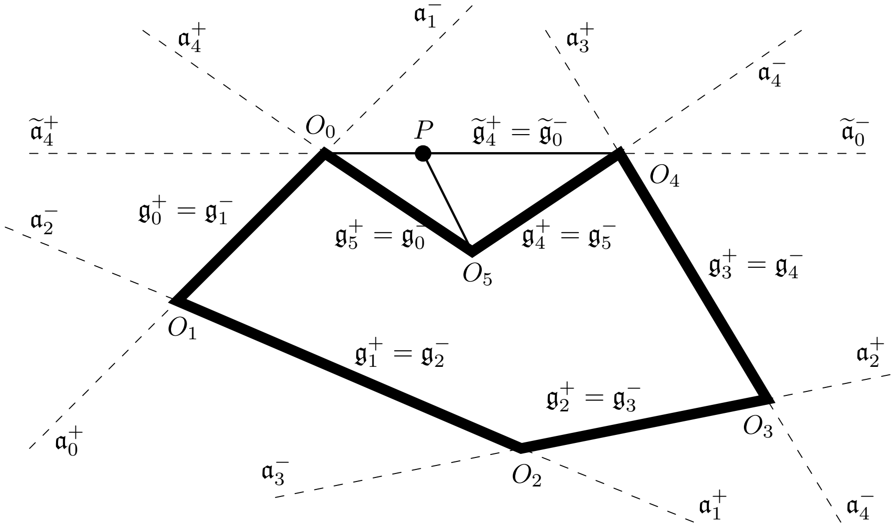

$\mathfrak{a}_4^+$
$\mathfrak{a}_1^-$
$\mathfrak{a}_3^+$
$\mathfrak{a}_4^-$
$\widetilde{\mathfrak{a}}_4^+$
O_0
P
$\widetilde{\mathfrak{g}}_4^+ = \widetilde{\mathfrak{g}}_0^-$
$\widetilde{\mathfrak{a}}_0^-$
O_4
$\mathfrak{a}_2^-$
$\mathfrak{g}_0^+ = \mathfrak{g}_1^-$
$\mathfrak{g}_5^+ = \mathfrak{g}_0^-$
$\mathfrak{g}_4^+ = \mathfrak{g}_5^-$
O_5
$\mathfrak{g}_3^+ = \mathfrak{g}_4^-$
O_1
$\mathfrak{g}_1^+ = \mathfrak{g}_2^-$
$\mathfrak{a}_2^+$
$\mathfrak{g}_2^+ = \mathfrak{g}_3^-$
O_3
$\mathfrak{a}_0^+$
$\mathfrak{a}_3^-$
O_2
$\mathfrak{a}_1^+$
$\mathfrak{a}_4^-$

FIG. 6. *Geometry of a nonconvex polygon.*

This representation is also used here for the analysis of radiation from the nonconvex polygon, but this time we say that $\mathfrak{D}_n$ is the maximal wedge which fits into the exterior of the polygon Γ and has a tip at the vertex O_n of Γ.

Obviously, in the case when Γ is convex, all of the newly defined wedges $\mathfrak{D}_n$ coincide with the corresponding exterior wedges of Γ, but in the case when Γ is nonconvex, this coincidence is no longer necessary. For example, in the configuration shown in Figure 6, the wedges $\mathfrak{D}_0$ and $\mathfrak{D}_4$ are smaller than the exterior wedges at the vertices O_0 and O_4.

To compute the field $U(r,\theta)$ that satisfies the Helmholtz equation in the exterior of Γ and has preassigned values on Γ we first assume that the values of $U(r,\theta)$ are already known on the segment O_0O_{N-2}, which can be considered as a side $\widetilde{\mathfrak{g}}_0^-$ of the convex polygon $\widetilde{\Gamma}$. Then, applying formula (5.10), we find that if (r,θ) is located outside $\widetilde{\Gamma}$, then the value of $U(r,\theta)$ can be represented as the mathematical expectation

$$\begin{aligned} U(r,\theta;\mu) = \mathrm{e}^{\mathrm{i}kr_\mu}\mathbf{E}\Big\{&\chi(\tau,t_1)f(\hat{\xi}_\tau,\eta_\tau;n_\tau)\mathrm{e}^{\mathrm{i}k[S(\tau)-\hat{\xi}_\tau]} \\ &+\chi_*(\tau,t_1)U(\hat{\xi}_{t_1},\eta_{t_1};n_{t_1})\mathrm{e}^{\mathrm{i}k[S(t_1)-\hat{\xi}_{t_1}]}\Big\}, \end{aligned} \tag{6.1}$$

where the averaging is extended over the stochastic processes $P_t = (\hat{\xi}_t,\eta_t)$, Λ_t, and n_t, which are controlled by the formulas (5.11)–(5.12) adapted to the polygon $\widetilde{\Gamma}$ and are stopped at the earlier of the exit times $t=\tau$ or $t=t_1$ defined, respectively, as the first times when P_t hits a side of the original polygon Γ or the side $\mathfrak{g}_0^- = O_0O_{N-2}$ of the auxiliary convex polygon $\widetilde{\Gamma}$.

The definition of the auxiliary exit time t_1 implies that P_{t_1} belongs to the segment O_0O_{N-2} which is located inside the wedge $\mathfrak{D}_{N-1}$ bounded by the half-lines $\mathfrak{a}_{N-1}^{\pm}$ and by the sides $\mathfrak{g}_{N-1}^{\pm}$ of the polygon Γ, where the values of $U(r,\theta)$ are assigned by the boundary conditions. This observation makes it possible to use the identities (2.4)–(2.5) and represent the unknown quantity $U(\hat{\xi}_{t_1},\eta_{t_1};n_{t_1})$ from the right-hand side of (6.1) as

$$U(\hat{\xi}_{t_1},\eta_{t_1};n_{t_1}) = U(r_{N-2},\theta_{N-2};N-2), \tag{6.2}$$

where r_{N-2} and θ_{N-2} are the $(N-2)$th coordinates computed by the formulas

$$r_{N-2} = \begin{cases} \mathcal{P}(\hat{\xi}_{t_1}, \eta_{t_1}, L_0^-) & \text{if} \quad n_{t_1-0} = j(N-1), \\ \mathcal{P}(\hat{\xi}_{t_1}, \alpha_{N-2} - \eta_{t_1}, L_0^-) & \text{if} \quad n_{t_1-0} = j(N-1) - 1, \end{cases} \tag{6.3}$$

and

$$\theta_{N-2} = \begin{cases} \alpha_{N-1} - \mathcal{F}(\hat{\xi}_{t_1}, \eta_{t_1}, L_0^-) & \text{if} \quad n_{t_1-0} = j(N-1) + 1, \\ \mathcal{F}(\hat{\xi}_{t_1}, \alpha_{N-2} - \eta_{t_1}, L_0^-) & \text{if} \quad n_{t_1-0} = j(N-1) - 1, \end{cases} \tag{6.4}$$

where j is an arbitrary integer. Then, evaluating the right-hand side of (6.2) by the formula (5.3) adjusted to the wedge $\mathfrak{D}_{N-1}$, we get a representation of $U(\hat{\xi}_{t_1}, \eta_{t_1}; n_{t_1})$ through yet unknown values of the field $U(r,\theta)$ on the half-lines $\mathfrak{a}_{N-1}^{\pm}$. Next, observing that both $\mathfrak{a}_{N-1}^{\pm}$ are located in the exterior of the convex polygon $\widetilde{\Gamma}$, we evaluate the new unknown quantities by another application of (6.1), and continuing recursively, we eventually come to the representation of $U(r,\theta)$ by the formulas

$$U(r,\theta;\mu) = \mathrm{e}^{\mathrm{i}kr_\mu} \mathbf{E} \left\{ f(\hat{\xi}_\tau, \eta_\tau; n_\tau) \mathrm{e}^{\mathrm{i}k[S(\tau) - \hat{\xi}_\tau]} \right\}, \qquad S(t) = \frac{1}{2} \int_0^\tau \hat{\xi}_t \mathrm{d}t - \Lambda_t, \tag{6.5}$$

where the stochastic processes n_t, $\hat{\xi}_t$, η_t, and Λ_t are defined by the transparent rules which are easier to understand than to formulate.

These processes start at the time $t = 0$ from initial positions defined by the formulas

$$\hat{\xi}_0 = r, \qquad \eta_0 = \theta, \qquad \Lambda_0 = 0, \qquad n_0 = \mu, \tag{6.6}$$

where, as in (5.11), the initial value $n_0 = \mu$ is not fixed but must be selected from the condition that $(r,\theta) \in \mathfrak{D}_\mu$. The further evolution of these processes is determined by their current position, which may be classified into the following distinct cases:

$$\begin{array}{lll} \text{Case } A_-: & P_t \in \widetilde{\mathfrak{a}}_{n_t}^-, & n_t = jN, \\ \text{Case } A_+: & P_t \in \widetilde{\mathfrak{a}}_{n_t}^+, & n_t = jN - 2, \\ \text{Case } B_-: & P_t \in \widetilde{\mathfrak{g}}_{n_t}^-, & n_t = jN, \\ \text{Case } B_+: & P_t \in \widetilde{\mathfrak{g}}_{n_t}^+, & n_t = jN - 2, \\ \text{Case } C_\pm: & P_t \in \mathfrak{a}_{n_t}^\pm, & \end{array}$$

where j may have any integer value. Then the processes n_t, Λ_t, and η_t are described by the following complicated but transparent rules:

$$n_{t+\mathrm{d}t} - n_t = \begin{cases} \pm 2 & \text{in Cases } A_\pm, \\ \pm 1 & \text{in Cases } B_\pm,\ C_\pm, \\ 0 & \text{otherwise}, \end{cases} \tag{6.7}$$

$$\Lambda_{t+\mathrm{d}t} - \Lambda_t = \begin{cases} |O_0 O_{N-2}| & \text{in Cases } A_\pm, \\ \rho_\pm & \text{in Cases } B_\pm, \\ L_{n_t}^\pm & \text{in Cases } C_\pm, \\ 0 & \text{otherwise}, \end{cases} \tag{6.8}$$

and

$$\eta_{t+\mathrm{d}t} - \eta_t = \mathrm{d}w_t^2 + \begin{cases} \widetilde{\alpha}_{n_t - 2} - \pi & \text{in Case } A_-, \\ \pi - \widetilde{\alpha}_{n_t} & \text{in Case } A_+, \\ \vartheta_\pm & \text{in Cases } B_\pm, \\ \alpha_{n_t - 1} - \pi & \text{in Case } C_-, \\ \pi - \alpha_{n_t} & \text{in Case } C_+, \\ 0 & \text{otherwise}, \end{cases} \tag{6.9}$$

where

$$\vartheta_- = \alpha_{N-1} - \mathcal{F}(\hat{\xi}_t, \eta_t, |O_0 O_{N-2}|), \qquad \vartheta_+ = \mathcal{F}(\hat{\xi}_t, \alpha_{N-1} - \eta_t, |O_0 O_{N-2}|), \tag{6.10}$$

$$\rho_- = \mathcal{P}(\hat{\xi}_t, \eta_t, |O_0 O_{N-2}|), \qquad \rho_+ = \mathcal{P}(\hat{\xi}_t, \alpha_{N-1} - \eta_t, |O_0 O_{N-2}|). \tag{6.11}$$

As for the radial motion $\hat{\xi}_t$, it is described by the same equations (5.14)–(5.16) as in the case of the convex polygon, but the term Λ_t involved there must get values from (6.8).

7. Solution of Problem-1. Results of the previous sections open the way to the solution of the fundamental Problem-1 identified in the end of section 3 as the principal part of the problem of diffraction by a polygon Γ.

Following the order established in the previous sections, we first consider the case when Γ is treated as a polygon with two infinite sides. More precisely, we start from the slightly more general problem formulated below, which includes nonvanishing boundary conditions on the faces of the wedge.

Problem-1w. Find a solution $U_\phi(r, \theta)$ of the Helmholtz equation that is bounded in a wedge $\mathfrak{O} : \alpha_1 < \theta < \alpha_2$, has the boundary values $U(r, \alpha_n) = f(r, \alpha_n)$ and the asymptote $\mathrm{e}^{-\mathrm{i}kr} U_\phi(r, \theta) = o(1)$ as $r \to \infty$, $\theta \neq \phi$, and satisfies the interface conditions

$$U_\phi(r, \phi + 0) - U_\phi(r, \phi - 0) = e^{\mathrm{i}kr}, \qquad \frac{\partial U_\phi(r, \phi + 0)}{\partial \theta} = \frac{\partial U_\phi(r, \phi - 0)}{\partial \theta}, \tag{7.1}$$

imposed on the ray $\theta = \phi$.

It is shown in [10] that to obtain the solution of this problem it suffices to split the wedge $\mathfrak{O}$ into two smaller wedges $\mathfrak{O}_-$ and $\mathfrak{O}_+$ defined by the inequalities $\alpha_1 < \theta < \phi$ and $\phi < \theta < \alpha_2$, respectively. Indeed, assuming that $U_\phi(r, \theta)$ is known on both sides of the ray $\theta = \phi$, we represent $U_\phi(r, \theta)$ in each of these wedges by formulas of the type (4.6). Then, matching these representations in order to enforce the interface conditions (7.1), we eventually arrive at the solution of the Problem-1w in the form

$$U_\phi(r, \theta) = \mathrm{e}^{\mathrm{i}kr} \mathbf{E} \left\{ f(\hat{\xi}_\tau, \eta_\tau) \mathrm{e}^{\mathrm{i}k[S(\tau) - \hat{\xi}_\tau]} + \sum_{\nu=1}^{t_\nu < \tau} \delta(t_\nu, \phi) \mathrm{e}^{\mathrm{i}kS(t_\nu)} \right\}, \quad S(t) = \frac{1}{2} \int_0^t \hat{\xi}_t \mathrm{d}t, \tag{7.2}$$

where $\hat{\xi}_t$ and η_t retain their meanings from (4.6); τ is the exit time defined as the first time when $\eta_t = \alpha_1$ or $\eta_t = \alpha_2$; the factor $\delta(t, \phi)$ is determined by the rule

$$\delta(t, \phi) = \begin{cases} 1 & \text{if} \quad \vartheta_{t-0} > \phi, \quad \vartheta_{t+0} < \phi, \\ -1 & \text{if} \quad \vartheta_{t-0} < \phi, \quad \vartheta_{t+0} > \phi, \\ 0 & \text{otherwise;} \end{cases} \tag{7.3}$$

and $\{t_n\}$, where $n \geq 1$, is a sequence of times when the angular motion η_t, running inside the interval $[\alpha_1, \alpha_2]$, touches the fixed point $\eta = \phi$.

With the solution of the auxiliary Problem-1w in hand we can now obtain the solution $U_{\mu,\phi}$ of Problem-1 for a nonconvex polygon Γ. Thus, assuming that this solution is already known on the entire faces of the wedge $\mathfrak{O}_\mu$, we can apply formula (7.2) and conclude that for any $(r, \theta) \in \mathfrak{O}_\mu$ the value of $U_{\mu,\phi}(r_\mu, \theta_\mu; \mu)$ can be evaluated as

$$U_{\mu,\phi}(r_\mu, \theta_\mu; \mu) = \mathrm{e}^{\mathrm{i}kr_\mu} \mathbf{E} \left\{ \sum_{\nu=1}^{t_\nu < \tau} \delta(t_\nu, \phi) \mathrm{e}^{\mathrm{i}kS(t_\nu)} + \chi(\hat{\xi}_\tau, \eta_\tau) U_{\mu,\phi}(\hat{\xi}_\tau, \eta_\tau; \mu) \mathrm{e}^{\mathrm{i}kS(\tau)} \right\}, \tag{7.4}$$

where

$$\chi(\hat{\xi},\eta) = \begin{cases} 0 & \text{if} \quad (\hat{\xi},\eta) \in \Gamma, \\ 1 & \text{otherwise.} \end{cases} \tag{7.5}$$

The right-hand side of (7.4) contains yet unknown values of the field $U_{\mu,\phi}(r_\mu, \theta_\mu; \mu)$ at the points $(\hat{\xi}_\tau, \eta_\tau)$ which belong to the boundary of the wedge $\mathfrak{O}_\mu$ but do not belong to the polygon Γ. These points, however, are located inside one of the wedges $\mathfrak{O}_{\mu\pm1}$, and therefore the field there can be computed by the formula (6.5) applied to the appropriate wedge $\mathfrak{O}_{\mu\pm1}$. Continuing the iterations, we eventually arrive at the expression

$$U_{\mu,\phi}(r,\theta) = \mathrm{e}^{\mathrm{i}kr_{\bar{\mu}}} \mathbf{E}\left\{\sum_{\nu=1}^{\infty} \delta(t_\nu, \phi; \mu) \mathrm{e}^{\mathrm{i}kS(t_\nu)}\right\}, \qquad S(t) = \frac{1}{2}\int_0^t \hat{\xi}_t \mathrm{d}t - \Lambda_t, \tag{7.6}$$

which is similar to (6.5)–(6.11) with the few exceptions described below.

Thus the mathematical expectation in (7.4) is computed over the stochastic processes n_t, $\hat{\xi}_t$, η_t, and Λ_t, which retain their meanings from (6.5)–(6.11), except that the integer-valued process n_t starts from the position (see Figure 7)

$$n_0 = \bar{\mu} = \begin{cases} \mu & \text{if} \quad 0 \le \theta_\mu \le \alpha_\mu, \\ \min\{m : \; P \in \mathfrak{O}_m\} & \text{if} \quad \theta_\mu > \alpha_\mu, \\ \max\{m : \; P \in \mathfrak{O}_m\} & \text{if} \quad \theta_\mu < 0, \end{cases} \tag{7.7}$$

which is the closest to the μ index of the wedge $\mathfrak{O}_{\bar{\mu}}$ containing the observation point $P = (r, \theta)$.

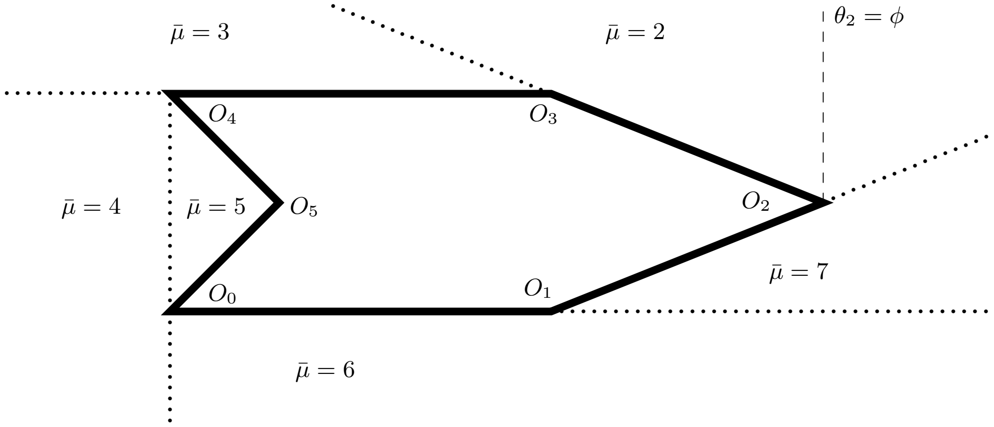
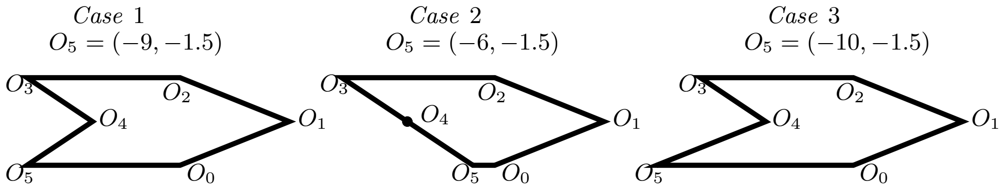


FIG. 7. *Selectionof $\bar{\mu}$ for $\theta_\mu > \phi$.*

Another feature of (7.4) which does not appear in (6.5) is the presence of the factor $\delta(t, \phi)$ determined by the rule

$$\delta(t, \phi; \mu) = \begin{cases} 1 & \text{if} \quad \phi < \eta_{\tau_\nu - 0}, \quad \eta_{\tau_\nu + 0} < \phi, \quad n_t = \mu, \\ -1 & \text{if} \quad \phi > \eta_{\tau_\nu - 0}, \quad \eta_{\tau_\nu + 0} > \phi, \quad n_t = \mu, \\ 0 & \text{otherwise,} \end{cases} \tag{7.8}$$



Fixed vertices: $O_0 = (-3.5, -1.5)$, $O_1 = (0, 0)$, $O_2 = (-3.5, 1.5)$, $O_3 = (-9, 1.5)$ $O_4 = (-6, 0)$

FIG. 8. *Samples of nonconvex polygons.*

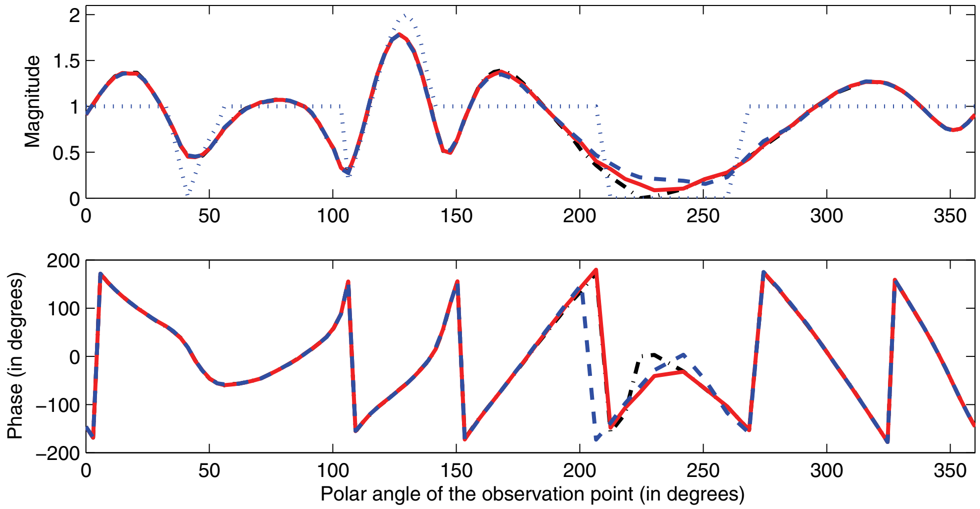


FIG. 9. *The total and geometric fields around the polygons from Figure* 8. *The wave fields were computed along circles of radius $R = 9$ with centers at the point $C = (-4.5, 0)$ in the middle of the polygons. In all of the diagrams the light dashed lines correspond to the geometric fields which comprise the incident and reflected plane waves, and the bold lines correspond to the total wave fields, which include the diffracted fields. The solid bold lines correspond to Case* 1, *the dash-dotted bold lines correspond to Case* 2, *and dashed bold lines correspond to Case* 3.

where ν enumerates the times $t = \tau_\nu$ when $n_t = \mu$ and the motion η_t touches the point $\eta = \phi$.

We don't provide here an extended derivation of (7.4)–(7.8) because it goes along the same lines as that discussed in section 6, and also because very similar formulas have actually been discussed in detail in the paper [8], devoted to diffraction by a convex polygon. The only difference between the solutions of Problem-1 obtained in that paper and here is that in [8] the analysis is based on the representation of the wave field in a wedge by the formulas (4.3)–(4.5), while the solution (7.4)–(7.8) is based on the more versatile formulas (4.6)–(4.9).

8. Example. To verify the feasibility of the obtained solution of the problem of diffraction by a nonconvex polygon, we conducted numerical simulations of the wave fields generated by an incident plane wave $U_* = \mathrm{e}^{-\mathrm{i}r\cos(\theta-\theta_*)}$, with $\theta_* = 60°$, from three short polygons shown in Figure 8. The vertices are located by standard Cartesian coordinates with the origin at O_1.

Figure 9 shows the magnitudes and the phases of the total and the geometric

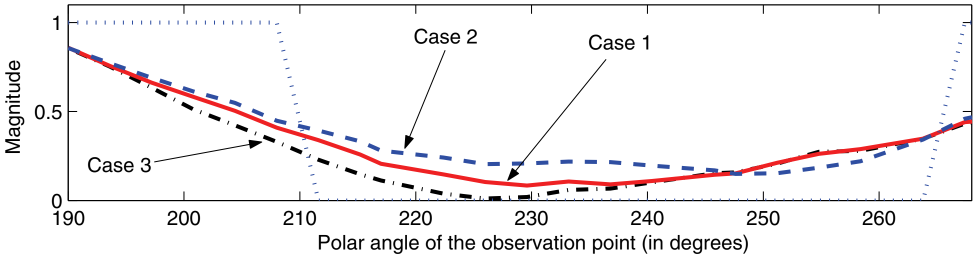


FIG. 10. *Zoomed view on magnitudes of the total fields from Figure* 9.

fields generated due to the scattering by the pentagons shown in Figure 8.

All of the numerical results were obtained by the approximation of the mathematical expectation (7.6) by the average of 1500 sample values of the functional depending on two standard one-dimensional Brownian motions w_t^1 and w_t^2. These Brownian motions are simulated by simple discreet random walks with jumps of distance $\Delta w = \pm\sqrt{\Delta t}$ following each other with the time increment $\Delta t = 0.01$. The standard deviation of the samples did not exceed the level $D = 0.5$, which suggests good convergence and stability of the solution. All computations were carried out using a MATLAB code of about 100 lines with no attempts at optimization. The code was run on a 900MHz notebook PC, and it required only a few seconds to calculate the results at each observation point.

The diagrams in Figure 9 clearly demonstrate the expected behavior of the simulated wave fields generated by the interaction of an incident plane wave with the polygon. Thus, in the absence of the scatter, the total wave field along the circles would have a unit magnitude and sinusoidal phase. In the presence of any of the scatterers from Figure 8, the laws of geometric optics predict a shadow zone around the ray $\theta = 235°$, and two zones illuminated by the waves reflected from the sides O_1O_2 and O_2O_3. It is also expected that the difference in the total fields should be most noticeable in the shadow zone where the magnitude of the total field should be highest in Case 2, medium in Case 1, and minimal in Case 3. These physically justified predictions are soundly confirmed by the presented graphs, which show that the total field in Case 3 almost vanishes around the ray $\theta \approx 225°$ located close to the vertex O_5 of the grounded polygon. At the same time the magnitude of the total field in Case 1 has a little variation in the domain between the rays $\theta = 220°$ and $\theta = 240°$, which is located on the approximately constant distance from the grounded polygon. These peculiarities of the wave fields are best seen in Figure 10, which reproduces in a larger scale the magnitudes of the total fields shown in Figure 9.

9. Conclusion. The probabilistic approach to wave propagation and diffraction made it possible to solve the problem of diffraction by a nonconvex polygon and, therefore, to expand the list of difficult problems of diffraction solved by this method.

In the past, the probabilistic random walk method led to explicit solutions of a number of nontrivial problems of diffraction, including the scalar problem of diffraction by a plane angular sector [6], the vector problem of diffraction of the electromagnetic wave by a wedge with anisotropic impedance faces [9], as well as problems of diffraction by a wedge with faces of variable impedance and by an arbitrary convex polygon. Although all of these problems are notoriously difficult for analysis by conventional methods, the obtained probabilistic solutions appear to be simple, transparent, compatible with intuitive ideas about diffraction, and easy for numerical implementation. The advantages of the random walk approach become even more apparent when it is applied to problems with several diffraction points. Thus, in [7, 8] we solved the problems of wave scattering in a half-plane with a piecewise constant impedance boundary condition, the problem of diffraction by a finite line segment with different impedance sides, and the problem of diffraction by an arbitrary convex polygon. Here we have taken the next step and extended the method to problems of diffraction by a nonconvex polygon. To make the presentation more transparent we limited ourselves to a perfectly reflecting nonconvex polygon of a specific shape, but, as shown in [10], the further extensions of the method to domains of virtually arbitrary shape with virtually arbitrary first-order impedance boundary conditions requires only routine modifications of the general method.

We hope that the presented results will stimulate further applications of the probabilistic methods and will extend the understanding of wave propagation and diffraction.

REFERENCES

[1] J. M. L. Bernard, *Scattering by a three-part impedance plane: A new spectral approach*, Quart. J. Mech. Appl. Math., 58 (2005), pp. 383–418.

[2] J. M. L. Bernard, *A spectral approach for scattering by impedance polygons*, Quart. J. Mech. Appl. Math., 59 (2006), pp. 517–550.

[3] V. A. Borovikov, *Diffraction by Polygons and Polyhedra*, Nauka, Moscow, 1966 (in Russian).

[4] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, eds., *Electromagnetic and Acoustic Scattering by Simple Shapes*, Hemisphere Publishing Corporation, New York, 1969.

[5] B. V. Budaev and D. B. Bogy, *Random walk approach to wave propagation in wedges and cones*, J. Acoust. Soc. Amer., 114 (2003), pp. 1733–1741.

[6] B. V. Budaev and D. B. Bogy, *Diffraction by a plane sector*, Proc. Roy. Soc. A, 460 (2004), pp. 3529–3546.

[7] B. V. Budaev and D. B. Bogy, *Two-dimensional problems of diffraction by finite collinear structures*, J. Acoust. Soc. Amer., 119 (2005), pp. 741–750.

[8] B. V. Budaev and D. B. Bogy, *Diffraction by a convex polygon with side-wise constant impedance*, Wave Motion, 43 (2006), pp. 631–645.

[9] B. V. Budaev and D. B. Bogy, *Diffraction of a plane electromagnetic wave by a wedge with anisotropic impedance faces*, IEEE Trans. Antennas Propagation, 54 (2006), pp. 1559–1567.

[10] B. V. Budaev and D. B. Bogy, *Novel solutions of the Helmholtz equation and their application to diffraction*, Proc. Roy. Soc. A, 463 (2007), pp. 1005–1027.

[11] E. B. Dynkin, *Markov Processes*, Grundlehren Math. Wiss. Einzeld. 121–122, Springer-Verlag, Berlin, 1965.

[12] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, Prentice–Hall Microwaves and Fields Series, Prentice–Hall, Englewood Cliffs, NJ, 1972.

[13] M. Freidlin, *Functional Integration and Partial Differential Equations*, Ann. Math. Stud. 109, Princeton University Press, Princeton, NJ, 1985.

[14] M. Idemen and A. Alkumru, *On a class of functional equations of the Wiener-Hopf type and their applications in n-part scattering problems*, IMA J. Appl. Math., 68 (2003), pp. 563–586.

[15] J. B. Keller, *A geometric theory of diffraction*, in Calculus of Variations and Its Applications, McGraw–Hill, New York, 1958, pp. 27–52.

[16] J. B. Keller, *Diffraction by polygonal cylinders*, in Electromagnetic Waves, University of Wisconsin Press, Madison, WI, 1962, pp. 129–137.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 419–452

MUTUALLY EXCLUSIVE SPIKY PATTERN AND SEGMENTATION MODELED BY THE FIVE-COMPONENT MEINHARDT–GIERER SYSTEM[*]

JUNCHENG WEI[†] AND MATTHIAS WINTER[‡]

Abstract. We consider the five-component Meinhardt–Gierer model for mutually exclusive patterns and segmentation, which was proposed in [H. Meinhardt and A. Gierer, *J. Theoret. Biol.*, 85 (1980), pp. 429–450]. We prove rigorous results on the existence and stability of mutually exclusive spikes which are located in different positions for the two activators. Sufficient conditions for existence and stability are derived, which depend in particular on the relative size of the various diffusion constants. Our main analytical methods are the Liapunov–Schmidt reduction and nonlocal eigenvalue problems. The analytical results are confirmed by numerical simulations.

Key words. pattern formation, mutual exclusion, stability, steady states

AMS subject classifications. Primary, 35B35, 92C15; Secondary, 35B40, 92D25

DOI. 10.1137/060673138

1. Introduction.

We analyze the five-component Meinhardt–Gierer system whose components are two activators and one inhibitor as well as two lateral activators. It has been introduced and very successfully used in various modeling aspects by Meinhardt and Gierer [11]. In particular, it can explain the phenomenon of mutual exclusion and can handle segmentation in the simplest case of two different segments. This model has been reviewed and its many implications discussed in detail by Meinhardt in Chapter 12 of [10].

The most important features of this system can be highlighted as *lateral activation of mutually exclusive states*. To each of the local activators a lateral activator is associated in a spatially nonlocal and time-delayed way. The consequence of the presence of the two lateral activators in the system is the possibility of having stable patterns which for the two activators are mutually exclusive; in other words, the patterns for the two activators are located in different positions. It is clear that mutually exclusive patterns are not possible for a three-component system with only two activators and one inhibitor since mutually exclusive patterns for the two activators could destabilize each other in various ways. Therefore the lateral activators are needed.

Numerical simulations of mutually exclusive patterns have been performed in [11], [10]. Many interesting features have been discovered and explained, but those works do not give analytical solutions, and they are not mathematically rigorous. To obtain mathematically rigorous results, in this study we show the existence and stability of mutually exclusive spikes in such a system.

The overall feedback mechanism of the system can be summarized as follows: *Lateral activation is coupled with self-activation and overall inhibition.* We will explain this in more detail after the system has been formulated quantitatively.

[*] Received by the editors October 24, 2006; accepted for publication (in revised form) July 17, 2008; published electronically November 5, 2008.
http://www.siam.org/journals/siap/69-2/67313.html

[†] Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong, China (wei@math.cuhk.edu.hk). The work of this author is supported by an Earmarked Grant of RGC of Hong Kong.

[‡] Department of Mathematical Sciences, Brunel University, Uxbridge UB8 3PH, United Kingdom (matthias.winter@brunel.ac.uk). The work of this author is supported by a BRIEF Award of Brunel University.

A widespread pattern in biology is *segmentation.* The mutual exclusion effect described in this paper is a special case of segmentation where only two different segments are present. Examples for biological segmentation are the body segments of insects or the segments of insect legs. The segments usually resemble each other strongly, but, on the other hand, they are different from each other. Segments may, for example, have an internal polarity which is often visible by bristles or hairs. This internal pattern within a segment depends on the position of the segment within the sequence in its natural state. In some biological cases a good understanding of how segment position and internal structure are related has been obtained. One famous example is surgical experiments on insects, e.g., cockroach legs. Creating a discontinuity in the normal neighborhood of structures by cutting a leg and pasting one piece to the end of another partial leg creates a discontinuity in the segment structure as some segments are missing their natural neighbors. By a process called intercalary regeneration new stable patterns in the cockroach leg are formed such that all segments get back their natural neighbors. However, the resulting pattern can be very different from any naturally occurring pattern.

For example, for cockroach legs, if the normal sequence of structures within a segment is 123 . . . 9, a combination of a partial leg 12345678 to which the piece 456789 is added first leads to the structure 12345678456789. Note the presence of the jump discontinuity in this sequence between the numbers 8 and 4. Now segment regulation adds the piece 765, which removes the discontinuity and leads to the final structure, 12345678**765**456789. This is different from the original natural structure, but nevertheless each segment has the same neighbors as in the natural situation.

In this example, which was experimentally verified by Bohn [1], it is not the natural sequence but the normal neighborhood which is regulated. It is exactly this neighboring structure which can be modeled mathematically using the system from [11] which is considered here, and this paper can be the starting point to a rigorous understanding of more complex segmentation phenomena.

Now we give a sociological application of mutual exclusion (see [11]). Consider two families. They can hardly live in exactly the same house, as this would lead to undesirable overcrowding. But if they live in the same street or neighborhood they can support, nurture, and benefit each other. Thus this collaborative behavior can lead to a rather stable situation. Indeed, stable coexisting states with concentration peaks remaining close but keeping a certain characteristic distance from each other are typical phenomena which are observed in quantitative models of systems modeling mutual exclusion, and they obviously resemble real-world behavior in this example very well.

This feedback mechanism of lateral activation coupled with overall inhibition can be quantified by formulating the effects of "activation," "lateral activation," and "inhibition" using the language of molecular reactions and invoking the law of mass action. Now we are going to discuss this in a quantitative manner. We will introduce the resulting model system first and then explain how these feedback mechanisms are represented by the terms in the model.

The original system from [11] (after rescaling and some simplifications) can be stated as follows:

$$\begin{cases} g_{1,t} = \epsilon^2 g_{1,xx} - g_1 + \dfrac{cs_2 g_1^2}{r}, \quad g_{2,t} = \epsilon^2 g_{2,xx} - g_2 + \dfrac{cs_1 g_2^2}{r}, \\ \tau r_t = D_r r_{xx} - r + cs_2 g_1^2 + cs_1 g_2^2, \\ \tau s_{1,t} = D_s s_{1,xx} - s_1 + g_1, \quad \tau s_{2,t} = D_s s_{2,xx} - s_2 + g_2. \end{cases} \tag{1.1}$$

Here $0 < \epsilon \ll 1$, $D_r > 0$ and $D_s > 0$ are diffusion constants, c is a positive reaction constant, and τ is a nonnegative time-relaxation constant (in [11] the choice $\tau = 1$ was made).

The x-indices indicate spatial derivatives. We will derive results for the system (1.1) on a bounded interval $\Omega = (-L, L)$ for $L > 0$ with Neumann boundary conditions. Some results for the system on the real line ($L = \infty$) will also be established and will be compared with the bounded interval case.

The first two components, the *activators* g_1 and g_2, activate themselves locally is due to the terms g_1^2 and g_2^2, respectively, in the first two equations.

The *lateral activators* are introduced in (1.1) by the fourth and fifth components s_1 and s_2 as follows: For both the activators, g_i, $i = 1, 2$, there are nonlocal and delayed versions s_i. Now s_1 acts as an activator to g_2, and s_2 acts as in activator to g_1 due to the terms s_2 in the first and s_1 in the second equation which have a positive feedback. The expression lateral activation is used since g_i activates g_{3-i} laterally through its nonlocal counterpart s_i rather than locally through g_i itself.

Lateral activation is finally coupled with overall inhibition as follows: The third component r acts as an *inhibitor* to both g_1 and g_2 due to the term r in the first and second equations, which has a negative feedback. Note also that both the local and the nonlocal activators have a positive feedback on r due to the terms $s_2 g_1^2$ and $s_1 g_2^2$ in the third equation.

This feedback mechanism is a generalization of the well-known Gierer–Meinhardt system [6] which has one local activator coupled to an inhibitor. We recall that the classical Gierer–Meinhardt system as well as the five-component system considered here are both Turing systems [13], as they allow spatial patterns to arise out of a homogeneous steady state by the so-called Turing instability. (Some analytical results for the existence and stability of a spiky Turing pattern for the Gierer–Meinhardt system have been obtained, for example, in [3], [4], [5], [9], [12], [14], [17], [18], [19].)

Now we state our rigorous results on the existence and stability of stationary, mutually exclusive, spiky patterns for the system (1.1).

We prove the *existence* of a spiky pattern with one spike for g_1 and one spike for g_2, which are located in different positions under the following conditions:

(i) the diffusivities of the two lateral activators are large compared to the inhibitor diffusivity and

(ii) the inhibitor diffusivity is large compared to the diffusivities of the two (local) activators.

We summarize the two main conditions (i) and (ii), which guarantee the existence of mutually exclusive spike patterns for (1.1), in the following:

$$\text{We assume that} \quad \epsilon^2 \ll C_1 D_r \le D_s \quad \text{for some constant } C_1 > 0. \tag{1.2}$$

We also prove the *stability* of these mutually exclusive spiky patterns, provided that certain conditions are met, which are of the type (1.2) with C_1 replaced by some new constant C_2.

In this paper we consider a pattern displaying one spike for g_1 and one for g_2 which are located in different positions.

In particular, we prove the existence of a mutually exclusive two-spike solution to the system (1.1) if $D_s/D_r > 4$. We show that this solution is stable if (i) $D_s/D_r > 43.33$ for $L = \infty$, or in general if (5.3) holds (condition for $O(1)$ eigenvalues), and if (ii) $D_s/D_r > 4$ (condition for $o(1)$ eigenvalues).

The main results will be stated in Theorem 3.2 on the existence of solutions and in Theorems 5.1 and 6.7 on the large and small eigenvalues, respectively, of the linearized problem at the solutions.

What do these results tell us about segmentation? As a first step, we have proved that in the case of two segments, which we call 1 and 2, the sequence 12 can exist and be stable, and we have found sufficient conditions for this effect to happen.

The case of $n > 2$ components will lead to a system with $2n+1$ components, which is very large and not easy to handle. Even in the case $n = 2$ for the five-component system investigated in this paper the analysis becomes rather lengthy. We expect that, following our approach, we will be able to prove existence and stability of n spikes in n different locations. We do not see any major obstacle, only that the proofs become more technical. We are currently working on this issue.

The outline of the paper is as follows. In section 2, we compute the amplitudes of the spikes for g_1 and g_2. In section 3, we determine the positions of the spikes and show the existence of steady states with mutually exclusive spikes. In section 4, we first derive the eigenvalue problem. Then we compute the large (i.e., $O(1)$) eigenvalues and derive sufficient conditions for the stability of solutions with respect to these. In section 5, we solve a nonlocal eigenvalue problem which has been delayed from section 4. In section 6, we give the most important steps and state the main result on the stability of solutions with respect to small (i.e., $o(1)$) eigenvalues. Sufficient conditions for this stability are derived. The technical details of the analysis of small eigenvalues is delayed to the appendices. Finally, in section 7, our results are confirmed by numerical simulations.

2. Computing the amplitudes.

We construct steady states of the form

$$g_1(x) = t_1 w\left(\frac{x - x_1}{\epsilon}\right)(1 + O(\epsilon)), \quad g_2(x) = t_2 w\left(\frac{x - x_2}{\epsilon}\right)(1 + O(\epsilon)),$$

where $w(y)$ is the unique positive and even homoclinic solution of the equation

$$w_{yy} - w + w^2 = 0 \tag{2.1}$$

on the real line decaying to zero at $\pm\infty$. Here we assume that the spikes for g_1 and g_2 have the same amplitude, i.e., $t_1 = t_2$. We often use different notation for the two amplitudes, as this will be important later when we consider stability, since there could be an instability which breaks the symmetry of having the same amplitudes. The analysis will show that t_1, t_2 and x_1, x_2 depend on ϵ but to leading order and after suitable scaling are independent of ϵ. To keep notation simple we will not explicitly indicate this dependence.

All functions used throughout this paper belong to the Hilbert space $H^2(-L, L)$, and the error terms are taken in the norm $H^2(-L, L)$ unless otherwise stated. After integrating (2.1), we get the relation

$$\int_R w(y)\,dy = \int_R w^2(y)\,dy, \tag{2.2}$$

which will be used frequently, often without explicitly stating it. We denote

$$w_1(x) = w\left(\frac{x - x_1}{\epsilon}\right), \quad w_2(x) = w\left(\frac{x - x_2}{\epsilon}\right). \tag{2.3}$$

Note that g_1 and g_2 are small-scale variables, as $\epsilon \ll 1$, and r, s_1, and s_2 are large-scale (with respect to the spatial variable). For steady states, using Green functions, these slow variables, to leading order, can be expressed by an integral representation.

To get this representation, g_1 in the last three equations of (1.1) can be expanded as

$$g_1(x) = t_1\epsilon\left(\int_R w\right)\delta_{x_1}(x) + O(\epsilon^2), \quad g_1^2(x) = t_1^2\epsilon\left(\int_R w^2\right)\delta_{x_1}(x) + O(\epsilon^2),$$

where $\delta_{x_1}(x) = \delta(x - x_1)$ is the Dirac delta distribution located at x_1. Similarly, for g_2 we have

$$g_2(x) = t_2\epsilon\left(\int_R w\right)\delta_{x_2}(x) + O(\epsilon^2), \quad g_2^2(x) = t_2^2\epsilon\left(\int_R w^2\right)\delta_{x_2}(x) + O(\epsilon^2).$$

Using the Green function $G_D(x,y)$, which is defined as the unique solution of the equation

(2.4)

$$D\Delta G_D(x,y) - G_D(x,y) + \delta_y(x) = 0, \quad -L < x < L, \quad G_{D,x}(-L,y) = G_{D,x}(L,y) = 0,$$

we can represent $s_1(x)$ by using the fourth equation of (1.1) as

$$s_1(x) = t_1\epsilon\left(\int_R w\right)G_{D_s}(x,x_1) + O(\epsilon^2). \tag{2.5}$$

An elementary calculation gives

$$G_D(x,y) = \begin{cases} \frac{\theta}{\sinh(2\theta L)}\cosh\theta(L+x)\cosh\theta(L-y), & -L < x < y < L, \\ \frac{\theta}{\sinh(2\theta L)}\cosh\theta(L-x)\cosh\theta(L+y), & -L < y < x < L, \end{cases} \tag{2.6}$$

with $\theta = 1/\sqrt{D}$. Note that

$$G_D(x,y) = \frac{1}{2\sqrt{D}}e^{-|x-y|/\sqrt{D}} - H_D(x,y), \tag{2.7}$$

where H_D is the regular part of the Green function G_D. In particular, for $L = \infty$, we have

$$G_D(x_1,x_2) = \frac{1}{2\sqrt{D}}e^{-|x-y|/\sqrt{D}} =: K_D(x_1,x_2). \tag{2.8}$$

In the same way, we derive

$$s_2(x) = t_2\epsilon\left(\int_R w\right)G_{D_s}(x,x_2) + O(\epsilon). \tag{2.9}$$

Now we compute the last two terms on the right-hand side (r.h.s.) of the third equation of (1.1) as follows:

$$\begin{aligned} cs_2g_1^2(x) &= cs_2(x_1)t_1^2\epsilon\left(\int_R w\right)\delta_{x_1}(x) + O(\epsilon^2) \\ &= ct_1^2t_2\epsilon^2\left(\int_R w\right)^2\delta_{x_1}(x)G_{D_s}(x_1,x_2) + O(\epsilon^3) \end{aligned}$$

and, similarly,

$$cs_1g_2^2(x) = ct_1t_2^2\epsilon^2\left(\int_R w\right)^2\delta_{x_2}(x)G_{D_s}(x_1,x_2) + O(\epsilon^3).$$

Now, using the third equation of (1.1), we can represent $r(x)$ by the Green function G_{D_r},

$$r(x) = ct_1t_2\epsilon^2\left(\int_R w\right)^2 G_{D_s}(x_1,x_2)(t_1G_{D_r}(x,x_1)+t_2G_{D_r}(x,x_2))+O(\epsilon^3). \tag{2.10}$$

Going back to the first equation in (1.1), we get

$$\epsilon^2\Delta g_1 - g_1 + \frac{cs_2g_1^2}{r} = t_1(\epsilon^2\Delta w_1 - w_1) + \frac{cs_2t_1^2w_1^2}{r} + O(\epsilon) = t_1\left[\frac{cs_2t_1}{r}-1\right]w_1^2 + O(\epsilon). \tag{2.11}$$

To have the same amplitudes of the two contributions in (2.11), we require

$$\frac{cs_2(x_1)t_1}{r(x_1)} = 1 + O(\epsilon). \tag{2.12}$$

Now we rewrite (2.12), using (2.9) and (2.10):

$$\frac{cs_2(x_1)t_1}{r(x_1)} = \frac{1}{\epsilon(\int_R w)(t_1G_{D_r}(x_1,x_1)+t_2G_{D_r}(x_1,x_2))} + O(\epsilon). \tag{2.13}$$

Thus, (2.12) for $x = x_1$ gives

$$t_1G_{D_r}(x_1,x_1)+t_2G_{D_r}(x_1,x_2) = \frac{1}{\epsilon\int_R w} + O(1). \tag{2.14}$$

In the same way, from the second equation in (1.1) we get

$$t_1G_{D_r}(x_1,x_2)+t_2G_{D_r}(x_2,x_2) = \frac{1}{\epsilon\int_R w} + O(1). \tag{2.15}$$

The relations (2.14), (2.15) are a linear system for the amplitudes t_1, t_2 of the spikes if their positions state that the amplitudes x_1, x_2 are known. Note that the amplitudes depend on the positions in leading order, as also the Green function G_{D_r} depends on its arguments in leading order. We say that the amplitudes are strongly coupled to the positions.

Note that the system (2.14), (2.15) has a unique solution t_1, t_2 since by (2.6)

$$G_{D_r}(x_1,x_1)G_{D_r}(x_2,x_2) - (G_{D_r}(x_1,x_2))^2 = \frac{\theta_r^2}{\sinh^2(2\theta_rL)}\cosh\theta_r(L-x_1)\cosh\theta_r(L+x_2)$$
$$\times[\cosh\theta_r(L+x_1)\cosh\theta_r(L-x_2) - \cosh\theta_r(L-x_1)\cosh\theta_r(L+x_2)] > 0$$

for $-L < x_2 < x_1 < L$, where $\theta_r = 1/\sqrt{D_r}$.

By symmetry, for $x_1 = -x_2$ we have $t_1 = t_2$. This is the case we are interested in. However, we have not yet shown that there are such positions x_1, x_2. This will be done in the next section.

For the special case $L = \infty$, we have $G_{D_r}(x_1,x_2) = \frac{1}{2\sqrt{D_r}}e^{-|x-y|/\sqrt{D_r}}$, and (2.14), (2.15) in this case are given by

$$t_1 + t_2e^{-|x_1-x_2|/\sqrt{D_r}} = \frac{2\sqrt{D_r}}{\epsilon\int_R w}, \quad t_2 + t_1e^{-|x_1-x_2|/\sqrt{D_r}} = \frac{2\sqrt{D_r}}{\epsilon\int_R w}.$$

Finally, we summarize the main result of this section as follows.

LEMMA 2.1. *Assume that $\epsilon > 0$ is small enough. Then for spike-solutions of* (1.1) *of the type*

$$g_1(x) = t_1 w\left(\frac{x - x_1}{\epsilon}\right)(1 + O(\epsilon)), \quad g_2(x) = t_2 w\left(\frac{x - x_2}{\epsilon}\right)(1 + O(\epsilon)),$$

where $w(y)$ is the unique positive and even solution of the equation

$$w_{yy} - w + w^2 = 0$$

on the real line decaying to zero at $\pm\infty$, the amplitudes t_1 and t_2 are given as the unique solution of the system

$$t_1 G_{D_r}(x_1, x_1) + t_2 G_{D_r}(x_1, x_2) = \frac{1}{\epsilon \int_R w} + O(1),$$
$$t_1 G_{D_r}(x_1, x_2) + t_2 G_{D_r}(x_2, x_2) = \frac{1}{\epsilon \int_R w} + O(1),$$

where G_D is the Green function defined in (2.4).

3. Existence of mutually exclusive spikes. In this section, we use the Liapunov–Schmidt reduction method to rigorously prove the existence of mutually exclusive spikes. We will get a sufficient condition on the locations of the spikes.

The problem here is that the linearization of the r.h.s. of the first equation in (1.1) around w_1 has an approximate nontrivial kernel. This comes from the fact that a derivative of (2.1) with respect to y gives

$$(w_y)_{yy} - w_y + 2ww_y = 0.$$

Thus, w_y belongs to the kernel of the linearization of (2.1) around w. Note that the function w_y represents the translation mode. Therefore a direct application of the implicit function theorem is not possible; one has to deal with this kernel first. This is the goal in this section.

Recall that for given $g_1, g_2 \in H_N^2(\Omega_\epsilon)$, where $\Omega_\epsilon = (-L/\epsilon, L/\epsilon)$ and $H_N^2(\Omega_\epsilon)$ denotes the space of all functions in $H^2(\Omega_\epsilon)$ satisfying the Neumann boundary condition, by the fourth equation of (1.1) s_1 is uniquely determined, by the fifth equation s_2 is uniquely determined, and finally by the third equation r is uniquely determined. Therefore, the steady state problem is reduced to solving the first two equations.

We are looking for solutions which satisfy

$$g_1(x) = t_1 w\left(\frac{x - x_1}{\epsilon}\right)(1 + O(\epsilon)), \quad g_2(x) = t_1 w\left(\frac{x + x_1}{\epsilon}\right)(1 + O(\epsilon))$$

with $g_1(x) = g_2(-x)$ $(x_1 > 0)$. By this reflection symmetry the problem is reduced to determining just one function: $g_1(x) = t_1 w_1(x) + v$.

We are now going to determine this function in *two steps.* Denoting the r.h.s. of the first equation of (1.1) by $S_\epsilon[t_1 w_1 + v]$, which is well defined for steady states, our problem can be written as follows: $S_\epsilon[t_1 w_1 + v] = 0$, where $S_\epsilon : H_N^2(\Omega_\epsilon) \to L^2(\Omega_\epsilon)$.

First step. Determine a small $v \in H^2(\Omega_\epsilon)$ with $\int_\Omega v \frac{dw_1}{dx}\, dx = 0$ such that

$$S_\epsilon[t_1 w_1 + v] = \beta\epsilon \frac{dw_1}{dx}. \tag{3.1}$$

Second step. Choose x_1 such that

$$\beta = 0. \tag{3.2}$$

We begin with the *first* step. To this end, we need to study the linearized operator

$$\tilde{L}_{\epsilon,x_1} : H^2(\Omega_\epsilon) \to L^2(\Omega_\epsilon) \quad \text{defined by} \quad \tilde{L}_{\epsilon,x_1} := S_\epsilon^{'}[t_1 w_1]\phi,$$

where $S_\epsilon^{'}[t_1 w_1]$ denotes the Fréchet derivative of the operator S_ϵ at $t_1 w_1$.

We define the approximate kernel and cokernel, respectively, as follows:

$$\mathcal{K}_{\epsilon,x_1} := \text{span}\left\{\epsilon \frac{dw_1}{dx}\right\} \subset H^2(\Omega_\epsilon), \quad \mathcal{C}_{\epsilon,x_1} := \text{span}\left\{\epsilon \frac{dw_1}{dx}\right\} \subset L^2(\Omega_\epsilon).$$

By projection, we define the operator

$$L_{\epsilon,x_1} = \pi^\perp_{\epsilon,x_1} \circ \tilde{L}_{\epsilon,x_1} : \mathcal{K}^\perp_{\epsilon,x_1} \to \mathcal{C}^\perp_{\epsilon,x_1},$$

where $\pi^\perp_{\epsilon,x_1}$ is the orthogonal projection in $L^2(\Omega_\epsilon)$ onto $\mathcal{C}^\perp_{\epsilon,x_1}$.

Then we have the following key result for the Liapunov–Schmidt reduction.

PROPOSITION 3.1. *There exist positive constants* $\bar{\epsilon}, \bar{\delta}, \lambda$ *such that we have for all* $\epsilon \in (0, \bar{\epsilon})$, $x_1 \in \Omega$ *with* $\min(|L + x_1|, |L - x_1|) > \bar{\delta}$,

$$\|L_{\epsilon,x_1}\phi\|_{L^2(\Omega_\epsilon)} \geq \lambda \|\phi\|_{H^2(\Omega_\epsilon)} \quad \textit{for all} \quad \phi \in \mathcal{K}^\perp_{\epsilon,x_1}. \tag{3.3}$$

Further, the map L_{ϵ,x_1} *is surjective.*

Proof of Proposition 3.1. We proceed by deriving a contradiction.

Suppose that (3.3) is false. Then there exist sequences $\{\epsilon_k\}, \{{x_1}^k\}, \{\phi^k\}$ with $\epsilon_k \to 0$, ${x_1}^k \in \Omega$, $\min(|L + x_1^k|, |L - x_1^k|) > \bar{\delta}$, $\phi^k = \phi_{\epsilon_k} \in K^\perp_{\epsilon_k, x_1^k}$, $k = 1, 2, \ldots,$ such that

$$\|L_{\epsilon_k, {x_1}^k}\phi^k\|_{L^2(\Omega_{\epsilon_k})} \to 0 \quad \text{as } k \to \infty, \quad \|\phi^k\|_{H^2(\Omega_{\epsilon_k})} = 1, \quad k = 1, 2, \ldots. \tag{3.4}$$

At first (after rescaling) ϕ_ϵ is defined only on Ω_ϵ. However, by a standard result (compare [7]) it can be extended to R such that its norm in $H^2(R)$ is still bounded by a constant independent of ϵ and x_1 for ϵ small enough. It is then a standard procedure to show that this extension converges strongly in $H^2(\Omega_\epsilon)$ to some limit ϕ_1 with $\|\phi_1\|_{L^2(R)} = 1$. For the details of the argument, we refer to [8].

The same analysis is performed for w_2 and its perturbation $\phi_{\epsilon,2}$. Then $\Phi = (\phi_1, \phi_2)^T$ solves the system

$$\begin{aligned} L_0\phi_1 - \frac{1}{\int_R w\, dy}\Bigg[& 2\hat{t}_1 G_{D_r}(x_1, x_1)\left(\int_R w\phi_1\, dy\right) + 2\hat{t}_1 G_{D_r}(x_1, x_2)\left(\int_R w\phi_2\, dy\right) \\ & + \hat{t}_2 G_{D_r}(x_1, x_2)\left(\int \phi_1\, dy\right) - \hat{t}_1 G_{D_r}(x_1, x_2)\left(\int \phi_2\, dy\right)\Bigg] = 0, \end{aligned} \tag{3.5}$$

$$\begin{aligned} L_0\phi_2 - \frac{1}{\int_R w\, dy}\Bigg[& 2\hat{t}_2 G_{D_r}(x_2, x_2)\left(\int w\phi_2\, dy\right) + 2\hat{t}_2 G_{D_r}(x_1, x_2)\left(\int w\phi_1\, dy\right) \\ & + \hat{t}_1 G_{D_r}(x_1, x_2)\left(\int_R \phi_2\, dy\right) - \hat{t}_2 G_{D_r}(x_1, x_2)\left(\int_R \phi_1\, dy\right)\Bigg] = 0, \end{aligned} \tag{3.6}$$

where $L_0\phi = \epsilon^2\phi_{yy} - \phi + 2w\phi$ and

$$\alpha_\epsilon = \left(\frac{1}{\epsilon\int_R w\,dy}\right) \quad \text{and} \quad \hat{t}_i = (\alpha_\epsilon)^{-1}t_i. \tag{3.7}$$

This system is the special case with $\lambda = 0$ of (4.7), (4.8) derived in section 4. To avoid doing this computation twice we have delayed it to section 4, where the more general case is considered.

Now, adding (3.5) and (3.6), we obtain

$$L_0(\phi_1+\phi_2) - w^2\left(\frac{2\int_R w(\phi_1+\phi_2)\,dy}{\int_R w^2\,dy}\right) = 0.$$

This implies by Theorem 1.4 of [15] that $\phi_1 = -\phi_2$, and, setting $\phi := \phi_1$, for ϕ we must have

$$L_0\phi - \frac{4}{4-c_0}\frac{w^2}{\int w^2\,dy}\int w\phi\,dy = \lambda\phi, \tag{3.8}$$

where $0 < c_0 < 2$ (compare (5.1) for $\lambda = 0$). Now by Theorem 1.4 of [15] we must have $\phi = 0$. This contradicts $\|\phi\|_{L^2(R)} = 1$. Therefore, (3.3) must be true.

By the closed range theorem it follows that the map L_{ϵ,x_1} is surjective. (The details are given, for example, in [8].) □

Based on this key result for the Liapunov–Schmidt reduction it is now fairly standard (see, for example, the works [8] and [16]) to derive that there exists a small $v \in H^2(\Omega_\epsilon)$ with $\int_\Omega v\frac{dw_1}{dx}\,dx = 0$ such that

$$S[t_1w_1 + v] = \beta\epsilon\frac{dw_1}{dx}.$$

This completes the *first* step.

We now turn to the *second* step. We have to show that $\beta = 0$ for a certain x_1. This amounts to showing that

$$\int_\Omega S[t_1w_1+v](x)\epsilon\frac{dw_1}{dx}\,dx = 0$$

for a certain x_1. Note that computing x_1 in fact means determining the locations of the spikes. To this end, we have to expand $S[t_1w_1+v](x_1+\epsilon y)$.

We compute

$$S[t_1w_1+v](x_1+\epsilon y) = t_1\left[\frac{cs_2(x_1+\epsilon y)t_1}{r(x_1+\epsilon y)} - 1\right]w_1^2(x_1+\epsilon y) + O(\epsilon^2).$$

Using (2.9), (2.10) and the expansions

$$G_D(x_1+\epsilon y, x_2) = G_D(x_1,x_2) + G_{D,x_1}(x_1,x_2)\epsilon y + O(\epsilon^2|y|^2)$$

and

$$G_D(x_1+\epsilon y, x_1) = G_D(x_1,x_1) - \frac{1}{2D}\epsilon|y| - \frac{1}{2}H_{D,x_1}(x_1,x_1)\epsilon y + O(\epsilon^2|y|^2),$$

where we have used (2.7), we get

$$\frac{cs_2(x_1+\epsilon y)t_1}{r(x_1+\epsilon y)} = \frac{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)}{G_{D_s}(x_1,-x_1)} \tag{3.9}$$

$$\times \frac{G_{D_s}(x_1,-x_1)+\frac{1}{2}G_{D_s,x_1}(x_1,-x_1)\epsilon y+O(\epsilon^2|y|^2)}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)-\epsilon|y|/(2D)+\frac{1}{2}(-H_{D_r,x_1}(x_1,x_1)+G_{D_r,x_1}(x_1,-x_1))\epsilon y}$$

$$= 1+\frac{G_{D_s,x_1}(x_1,-x_1)}{2G_{D_s}(x_1,-x_1)}\epsilon y - \frac{G_{D_r,x_1}(x_1,-x_1)-H_{D_r,x_1}(x_1,x_1)}{2[G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)]}\epsilon y + O(\epsilon^2 y^2)$$
$$+\,\text{even term in } y.$$

This implies

$$\int_\Omega S[w_1+v](x)\epsilon\frac{dw_1}{dx}\,dx$$

$$= \frac{1}{2}\left[\frac{G_{D_s,x_1}(x_1,-x_1)}{G_{D_s}(x_1,-x_1)} - \frac{G_{D_r,x_1}(x_1,-x_1)-H_{D_r,x_1}(x_1,x_1)}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)}\right]\epsilon y\int_R yw^2\frac{dw}{dy}\,dy+\epsilon^2 W_\epsilon(x_1),$$

where $W_\epsilon(x_1)=O(\epsilon)$, uniformly for $0\le x_1\le L$.

Using (2.6), we further compute

$$F(x_1) := \frac{G_{D_s,x_1}(x_1,-x_1)}{G_{D_s}(x_1,-x_1)} - \frac{G_{D_r,x_1}(x_1,-x_1)-H_{D_r,x_1}(x_1,x_1)}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)}$$

$$= -\theta_s\frac{\sinh 2\theta_s(L-x_1)}{\cosh^2\theta_s(L-x_1)} - \theta_r\frac{\sinh 2\theta_r x_1-\sinh 2\theta_r(L-x_1)}{\cosh\theta_r(L-x_1)[\cosh\theta_r(L-x_1)+\cosh\theta_r(L+x_1)]},$$

where $\theta = 1/\sqrt{D}$. We have to determine x_1 such that $F(x_1)=0$. Note that

$$F(0) = -\theta_s\frac{\sinh 2\theta_s L}{\cosh^2\theta_s L}+\theta_r\frac{\sinh 2\theta_r L}{2\cosh^2\theta_r L} > 0$$

if

$$\frac{\theta_s}{\theta_r} < \frac{1}{2}\frac{\tanh\theta_r L}{\tanh\theta_s L}. \tag{3.10}$$

The inequality (3.10) is satisfied if, for fixed L, θ_r is large compared to θ_s.

In the limit $L\to 0$ the condition (3.10) converges to $\frac{\theta_s}{\theta_r} < 1/\sqrt{2}$. In the limit $L\to\infty$, (3.10) gives $\frac{\theta_s}{\theta_r} < 1/2$. For general $L\in(0,\infty)$ we can write (3.10) as follows: $\frac{\theta_s}{\theta_r} < \alpha(L)$ with $\frac{1}{2} < \alpha(L) < \frac{1}{\sqrt{2}}$.

Going back to the original diffusion constants, the inequality (3.10) is equivalent to

$$\frac{D_s}{D_r} > 4\frac{\tanh^2\theta_s L}{\tanh^2\theta_r L}. \tag{3.11}$$

In the limit $L\to 0$, (3.11) gives $\frac{D_s}{D_r} > 2$ and, in the limit $L\to\infty$, we can write (3.11) as follows: $\frac{D_s}{D_r} > 4$.

For all $L \in (0, \infty)$ we can write (3.11) as follows: $\frac{D_s}{D_r} > \beta(L)$ for some continuous function $\beta(L) \in (2, 4)$.

Note that (3.11) holds if

$$\frac{D_s}{D_r} > 4. \tag{3.12}$$

This is not the optimal condition, but it is rather handy and easy to check.

On the other hand,

$$F(L/2) = -\theta_s \frac{\sinh \theta_s L}{\cosh^2(\theta_s L/2)} < 0.$$

By the intermediate value theorem, under the condition (3.11), there exists an $x_1 \in (0, L/2)$ such that $F(x_1) = 0$. There exists no such $x_1 \in [L/2, L)$, since the function F is negative in that interval.

Note that $F(L/2) \to 0$ as $\theta_s \to 0$. This implies that $x_1 \to L/2$ as $\theta_s \to 0$.

We now show that the zero $x_1 \in [0, L/2]$ of F is unique by proving that $F'(x_1) < 0$ for $x_1 \in (0, L/2)$ if

$$\frac{\theta_s}{\theta_r} < \frac{\tanh(\theta_r L/2)}{\sqrt{2}\tanh(\theta_s L/2)}. \tag{3.13}$$

We compute

$$F'(x_1) = 2\theta_s^2 \frac{1}{\cosh^2 \theta_s(L - x_1)} - \theta_r^2 \frac{1}{\cosh^2 \theta_r(L - x_1)}$$

$$- \theta_r^2 \frac{[\cosh \theta_r(L - x_1) + \cosh \theta_r(L + x_1)]^2 - [\sinh \theta_r(L - x_1) + \sinh \theta_r(L + x_1)]^2}{[\cosh \theta_r(L - x_1) + \cosh \theta_r(L + x_1)]^2}.$$

Therefore, taking into consideration only the first two terms and noting that the last term is negative, we have $F'(x_1) < 0$ if (3.13) holds, and in this case, the solution for x_1 is unique.

Note that (3.13) holds if $\frac{\theta_s}{\theta_r} < \frac{1}{\sqrt{2}}$ or, equivalently, $\frac{D_s}{D_r} > 2$.

Therefore (3.10) and (3.13) are both true if $\frac{\theta_s}{\theta_r} < \frac{1}{2}$ or, equivalently, $\frac{D_s}{D_r} > 4$.

Now for (3.13), since $F'(x_1) \neq 0$, a standard degree argument shows that for $\epsilon \ll 1$ there exists a unique x_1^ϵ depending on ϵ such that $\int_\Omega S[w_1 + v](x)\epsilon \frac{dw_1}{dx}\, dx = 0$. Further, $x_1^\epsilon \to x_1$ as $\epsilon \to 0$, where x_1 satisfies

$$\frac{G_{D_s, x_1}(x_1, -x_1)}{G_{D_s}(x_1, -x_1)} - \frac{G_{D_r, x_1}(x_1, -x_1) - H_{D_r, x_1}(x_1, x_1)}{G_{D_r}(x_1, x_1) + G_{D_r}(x_1, -x_1)} = 0.$$

Thus we have shown existence and at the same time located the positions of the spikes. We summarize this result in the following theorem.

Theorem 3.2. *There exist mutually exclusive, spiky steady states to* (1.1) *in* $(-L, L)$ *with Neumann boundary conditions such that*

$$g_1^\epsilon(x) = t_1^\epsilon w\left(\frac{x - x_1^\epsilon}{\epsilon}\right)(1 + O(\epsilon)), \quad g_2^\epsilon(x) = t_1^\epsilon w\left(\frac{x + x_1^\epsilon}{\epsilon}\right)(1 + O(\epsilon)) \tag{3.14}$$

with

$$t_1^\epsilon = \frac{1}{\epsilon \int_R w\, dy\, (G_{D_r}(x_1, x_1) + G_{D_r}(x_1, -x_1))} + O(1) \tag{3.15}$$

and $x_1^\epsilon \to x_1$ *as* $\epsilon \to 0$*, where*

$$\frac{G_{D_s,x_1}(x_1,-x_1)}{G_{D_s}(x_1,-x_1)} - \frac{G_{D_r,x_1}(x_1,-x_1) - H_{D_r,x_1}(x_1,x_1)}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,-x_1)} = 0. \tag{3.16}$$

If $D_s/D_r > 4$*, then* (3.16) *has a unique solution* $x_1 \in (0,L/2]$ *and no solution in* $(L/2,L]$*. Further,* $x_1 \to L/2$ *as* $\theta_s \to 0$*.*

Finally, we compute the equation for x_1 in the limit $L \to \infty$. In this limit, x_1 satisfies

$$\frac{\theta_s}{\theta_r} = \frac{e^{-2\theta_r x_1}}{1+e^{-2\theta_r x_1}} + O(e^{-CL})$$

for some $C > 0$ independent of x_1. This is equivalent to

$$e^{2|x_1|/\sqrt{D_r}} = \sqrt{\frac{D_s}{D_r}} - 1 + O(e^{-CL}). \tag{3.17}$$

This concludes our study of existence. In the following sections we consider the stability issue.

4. Stability I: The eigenvalue problem and the large eigenvalues. Now we study the (linearized) stability of this mutually exclusive steady state. To this end, we first derive the linearized operator around the steady state $(g_1^\epsilon, g_2^\epsilon, r^\epsilon, s_1^\epsilon, s_2^\epsilon)$ given in Theorem 3.2.

We perturb the steady state as follows:

$$g_1 = g_1^\epsilon + \phi_1^\epsilon e^{\lambda t}, \quad g_2 = g_2^\epsilon + \phi_2^\epsilon e^{\lambda t}, \quad r = r^\epsilon + \psi^\epsilon e^{\lambda t},$$

$$s_1 = s_1^\epsilon + \eta_1^\epsilon e^{\lambda t}, \quad s_2 = s_2^\epsilon + \eta_2^\epsilon e^{\lambda t}.$$

By linearization we obtain the following eigenvalue problem (dropping superscripts ϵ):

$$\begin{cases} \lambda_\epsilon \phi_1 = \epsilon^2 \phi_{1,xx} - \phi_1 + \dfrac{c\eta_2 g_1^2}{r} + \dfrac{2cs_2 g_1 \phi_1}{r} - \dfrac{cs_2 g_1^2 \psi}{r^2}, \\ \lambda_\epsilon \phi_2 = \epsilon^2 \phi_{2,xx} - \phi_2 + \dfrac{c\eta_1 g_2^2}{r} + \dfrac{2cs_1 g_2 \phi_2}{r} - \dfrac{cs_1 g_2^2 \psi}{r^2}, \\ \tau\lambda_\epsilon \psi = D_r \psi_{xx} - \psi + c\eta_2 g_1^2 + 2cs_2 g_1 \phi_1 + c\eta_1 g_2^2 + 2cs_1 g_2 \phi_2, \\ \tau\lambda_\epsilon \eta_1 = D_s \eta_{1,xx} - \eta_1 + \phi_1, \\ \tau\lambda_\epsilon \eta_2 = D_s \eta_{2,xx} - \eta_2 + \phi_2, \end{cases} \tag{4.1}$$

where all components belong to the space $H_N^2(\Omega)$.

We now analyze the case $\lambda_\epsilon \to \lambda_0 \neq 0$ (large eigenvalues). After rescaling and taking the limit $\epsilon \to 0$ in (4.1) and noting that ϕ_i converges locally in $H^2(-L/\epsilon, L/\epsilon)$, we get for the first two components, using the approximations of g_1 and g_2 given in Theorem 3.2:

$$\epsilon^2 \Delta \phi_1 - \phi_1 + \frac{2cs_2(x_1)t_1 w_1 \phi_1}{r(x_1)} - \frac{cs_2(x_1)t_1^2 w_1^2}{r^2(x_1)}\psi(x_1) + \frac{c\eta_2(x_1)t_1^2 w_1^2}{r(x_1)} = \lambda\phi_1, \tag{4.2}$$

$$\epsilon^2 \Delta \phi_2 - \phi_2 + \frac{2cs_1(x_2)t_2 w_2 \phi_2}{r(x_2)} - \frac{cs_1(x_2)t_2^2 w_2^2}{r^2(x_2)}\psi(x_2) + \frac{c\eta_2(x_2)t_2^2 w_2^2}{r(x_2)} = \lambda\phi_1. \tag{4.3}$$

Now, in (4.2) and (4.3) we calculate the terms $\psi(x)$ and $\eta_1(x)$ and $\eta_2(x)$, respectively. To get $\psi(x)$, using the Green function G_{D_r}, we solve the linear equation for ψ given by

$$D_r\psi_{xx} - \psi + 2cs_2t_1w_1\phi_1 + 2cs_1t_2w_2\phi_2 + c\eta_2t_1^2w_1^2 + c\eta_1t_2^2w_2^2 = 0,$$

where again for g_1 and g_2 we have used the asymptotic expansions of Theorem 3.2. For simplicity, we study the case $\tau = 0$. Then the stability result extends to small τ as well, since we know that $|\lambda_\epsilon| \le C$ for all eigenvalues such that $\lambda_\epsilon > -c_0$ for some small $c_0 > 0$, which can be shown by a simple argument based on quadratic forms. This gives

$$\begin{aligned}\psi(x) \sim &\left[2cs_2(x_1)t_1\epsilon\left(\int_R w\phi_1\,dy\right) + c\eta_2(x_1)t_1^2\epsilon\int_R w^2\,dy\right]G_{D_r}(x,x_1)\\ &+\left[2cs_1(x_2)t_2\epsilon\left(\int_R w\phi_2\,dy\right) + c\eta_1(x_2)t_2^2\epsilon\int_R w^2\,dy\right]G_{D_r}(x,x_2).\end{aligned} \tag{4.4}$$

Similarly, using G_{D_s}, we compute

$$\eta_1(x) \sim \epsilon G_{D_s}(x,x_1)\int_R \phi_1\,dy, \quad \eta_2(x) \sim \epsilon G_{D_s}(x,x_2)\int_R \phi_2\,dy. \tag{4.5}$$

Recalling from (2.5) and (2.9) that

$$s_1(x) \sim \epsilon t_1\left(\int_R w\,dy\right)G_{D_s}(x,x_1), \quad s_2(x) \sim \epsilon t_2\left(\int_R w\,dy\right)G_{D_s}(x,x_2),$$

we get from (4.4)

$$\begin{aligned}\psi(x)\sim&\left[2ct_1t_2\epsilon^2\left(\int_R w\,dy\right)\left(\int_R w\phi_1\,dy\right) + ct_1^2\epsilon^2\left(\int_R w\,dy\right)\int_R \phi_2\,dy\right]G_{D_s}(x_1,x_2)G_{D_r}(x,x_1)\\ &+\left[2ct_1t_2\epsilon^2\left(\int_R w\,dy\right)\left(\int_R w\phi_2\,dy\right) + ct_2^2\epsilon^2\left(\int_R w\,dy\right)\int_R \phi_1\,dy\right]G_{D_s}(x_1,x_2)G_{D_r}(x,x_2).\end{aligned} \tag{4.6}$$

Further, recall from (2.10) that

$$r(x) = ct_1t_2\epsilon^2\left(\int_R w\,dy\right)^2 G_{D_s}(x_1,x_2)(t_1G_{D_r}(x,x_1) + t_2G_{D_r}(x,x_2)) + O(\epsilon^3).$$

Substituting into (4.2), we get for the coefficient of $\int_R \phi_1\,dy$ on the r.h.s.

$$\begin{aligned}&-\frac{cs_2(x_1)t_1^2w_1^2}{r^2(x_1)}c\epsilon^2\left(\int_R w\,dy\right)t_2^2G_{D_s}(x_1,x_2)G_{D_r}(x_1,x_2) + O(\epsilon^2)\\ &= -\frac{w_1^2}{s_2(x_1)}\epsilon^2\left(\int_R w\,dy\right)t_2^2G_{D_s}(x_1,x_2)G_{D_r}(x_1,x_2) + O(\epsilon^2)\\ &= -\epsilon t_2w_1^2G_{D_r}(x_1,x_2) + O(\epsilon^2).\end{aligned}$$

Similarly, the coefficient for $\int_R \phi_2\,dy$ is calculated as

$$
\begin{aligned}
&-\frac{cs_2(x_1)t_1^2w_1^2}{r^2(x_1)}c\epsilon^2\left(\int_R w^2\,dy\right)t_1^2G_{D_s}(x_1,x_2)G_{D_r}(x_1,x_1)+\frac{c\epsilon G_{D_s}(x_1,x_2)t_1^2w_1^2}{r(x_1)}+O(\epsilon^2)\\
&=-\frac{w_1^2}{s_2(x_1)}\epsilon^2\left(\int_R w^2\,dy\right)t_1^2G_{D_s}(x_1,x_2)G_{D_r}(x_1,x_1)+\frac{w_1^2}{s_2(x_1)}\epsilon t_1G_{D_s}(x_1,x_2)+O(\epsilon^2)\\
&\quad=-\frac{\epsilon t_1^2w_1^2}{t_2}G_{D_r}(x_1,x_1)+\frac{t_1}{t_2\int_R w\,dy}w_1^2+O(\epsilon^2)=\epsilon t_1w_1^2G_{D_r}(x_1,x_2)+O(\epsilon^2).
\end{aligned}
$$

Here we have used (2.14). Then (4.2) gives the nonlocal eigenvalue problem (NLEP)

$$
\begin{aligned}
&L_0\phi_1-\frac{1}{\int_R w\,dy}\Bigg[2\hat{t}_1G_{D_r}(x_1,x_1)\left(\int_R w\phi_1\,dy\right)+2\hat{t}_1G_{D_r}(x_1,x_2)\left(\int_R w\phi_2\,dy\right)\\
&\qquad+\hat{t}_2G_{D_r}(x_1,x_2)\left(\int_R \phi_1\,dy\right)-\hat{t}_1G_{D_r}(x_1,x_2)\left(\int_R \phi_2\,dy\right)\Bigg]=\lambda\phi_1,
\end{aligned}
\tag{4.7}
$$

where $L_0\phi=\epsilon^2\phi_{yy}-\phi+2w\phi$ and $\hat{t}_i$ has been defined in (3.7). In the same way, for (4.3) we obtain

$$
\begin{aligned}
&L_0\phi_2-\frac{1}{\int_R w\,dy}\Bigg[2\hat{t}_2G_{D_r}(x_2,x_2)\left(\int_R w\phi_2\,dy\right)+2\hat{t}_2G_{D_r}(x_1,x_2)\left(\int_R w\phi_1\,dy\right)\\
&\qquad+\hat{t}_1G_{D_r}(x_1,x_2)\left(\int_R \phi_2\,dy\right)-\hat{t}_2G_{D_r}(x_1,x_2)\left(\int_R \phi_1\,dy\right)\Bigg]=\lambda\phi_2,
\end{aligned}
\tag{4.8}
$$

where $\phi_1,\phi_2\in H^2(R)$. Set $\phi=(\phi_1,\phi_2)$ and denote by $L\phi$ the left-hand sides (l.h.s.) of (4.7) and (4.8), respectively.

Then, writing (4.7) and (4.8) in matrix notation, we have following the vectorial NLEP:

$$
L\phi=\Delta\phi-\phi+2w\phi-\left[\mathcal{B}\int_R \phi\,dy+2\mathcal{C}\left(\int_R w\phi\,dy\right)\right]\left(\int_R w\,dy\right)^{-1}w^2,
$$

where

$$
\mathcal{B}=G_{D_r}(x_1,x_2)\begin{pmatrix}\hat{t}_2 & -\hat{t}_1\\ -\hat{t}_2 & \hat{t}_1\end{pmatrix}=\frac{G_{D_r}(x_1,x_2)}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,x_2)}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}
\tag{4.9}
$$

and

$$
\begin{aligned}
\mathcal{C}&=\begin{pmatrix}\hat{t}_1G_{D_r}(x_1,x_1) & \hat{t}_1G_{D_r}(x_1,x_2)\\ \hat{t}_2G_{D_r}(x_1,x_2) & \hat{t}_2G_{D_r}(x_2,x_2)\end{pmatrix}\\
&=\frac{1}{G_{D_r}(x_1,x_1)+G_{D_r}(x_1,x_2)}\begin{pmatrix}G_{D_r}(x_1,x_1) & G_{D_r}(x_1,x_2)\\ G_{D_r}(x_1,x_2) & G_{D_r}(x_2,x_2)\end{pmatrix}.
\end{aligned}
\tag{4.10}
$$

Here we have used that (2.14), (2.15) imply

$$
\hat{t}_1G_{D_r}(x_1,x_1)+\hat{t}_2G_{D_r}(x_1,x_2)=1,\qquad \hat{t}_1G_{D_r}(x_1,x_2)+\hat{t}_2G_{D_r}(x_2,x_2)=1
\tag{4.11}
$$

and therefore

$$\hat{t}_i = \frac{G_{D_r}(x_{3-i}, x_{3-i}) - G_{D_r}(x_1, x_2)}{G_{D_r}(x_1, x_1)G_{D_r}(x_2, x_2) - (G_{D_r}(x_1, x_2))^2}, \quad i = 1, 2. \tag{4.12}$$

In the special case when $G_{D_r}(x_1, x_1) = G_{D_r}(x_2, x_2)$ we have

$$\hat{t}_1 = \hat{t}_2 = \frac{1}{G_{D_r}(x_1, x_1) + G_{D_r}(x_1, x_2)}. \tag{4.13}$$

Now, adding (4.7) and (4.8), we obtain

$$L_0(\phi_1 + \phi_2) - w^2 \left(\frac{2\int_R w(\phi_1 + \phi_2)\, dy}{\int_R w^2\, dy} \right) = \lambda(\phi_1 + \phi_2),$$

which implies by Theorem 1.4 of [15] that $\phi_1 + \phi_2 = 0$ if $\mathrm{Re}(\lambda_0) \geq 0$. So we set $\phi_2 = -\phi_1 = -\phi$.

From (4.7), we obtain a scalar NLEP for ϕ,

$$L_0\phi - \frac{w^2}{\int_R w^2\, dy} \left[c_0 \int_R w\phi\, dy + d_0 \int_R \phi\, dy \right] = \lambda\phi, \tag{4.14}$$

where

$$c_0 = \frac{2(G_{D_r}(x_1, x_1) - G_{D_r}(x_1, x_2))}{G_{D_r}(x_1, x_1) + G_{D_r}(x_1, x_2)}, \quad d_0 = \frac{2G_{D_r}(x_1, x_2)}{G_{D_r}(x_1, x_1) + G_{D_r}(x_1, x_2)}. \tag{4.15}$$

Note that $0 < c_0 < 2$ and $0 < d_0 < 1$.

In the following section we study the NLEP (4.14). It determines the stability or instability of the large eigenvalues of (4.1) if $0 < \epsilon < \epsilon_0$ for a suitably chosen ϵ_0. By our analysis, instabilities for small $\epsilon > 0$ imply instabilities for $\epsilon = 0$. On the other hand, by an argument of Dancer [2], an instability for $\epsilon = 0$ also gives an instability for small $\epsilon > 0$.

Note that the NLEP problem here is quite different from those studied in [4], [5], [14], and [15].

In the next section we study this eigenvalue problem and complete the investigation of $O(1)$ eigenvalues for (4.1).

5. Stability II: A nonlocal eigenvalue problem. In this section, we study the NLEP (4.14) to determine whether or not there are large eigenvalues, i.e., eigenvalues of the order $O(1)$ as $\epsilon \to 0$, which destabilize the mutually exclusive spiky pattern. Integrating (4.14), we have

$$\int_R \phi\, dy = \frac{2 - c_0}{\lambda + 1 + d_0} \int_R w\phi\, dy.$$

Substituting this back into (4.14), we can eliminate the term $\int_R \phi\, dy$. This gives

$$L_0\phi - \mu(\lambda) \frac{w^2}{\int_R w^2\, dy} \int_R w\phi\, dy = \lambda\phi, \quad \text{where} \quad \mu(\lambda) = \frac{c_0\lambda + 2}{\lambda + 2 - c_0/2}. \tag{5.1}$$

Here we have used that $c_0 + 2d_0 = 2$. Applying inequality (2.22) of [18], we get

$$\frac{\int_R w^3\, dy}{\int_R w^2\, dy} |\mu(\lambda_0) - 1|^2 + \mathrm{Re}(\overline{\lambda_0}(\mu(\lambda_0) - 1)) \leq 0 \quad \text{if } \mathrm{Re}(\lambda_0) \geq 0. \tag{5.2}$$

Observe that after multiplying (2.1) by w and by w', respectively, and integrating, we get

$$\int_R w^3\,dy = \frac{6}{5}\int_R w^2\,dy.$$

So, assuming without loss of generality that $\lambda_0 = +\sqrt{-1}\lambda_I$, we get for the l.h.s. in (5.2) the following:

$$\frac{6}{5}\left|\frac{c_0\lambda_0+2}{\lambda_0+1+d_0}-1\right|^2 + \mathrm{Re}\left(\overline{\lambda_0}\left(\frac{c_0\lambda_0+2}{\lambda_0+1+d_0}-1\right)\right)$$

$$= \frac{6}{5}\frac{(c_0-1)^2|\lambda_0|^2+(1-d_0)^2}{|\lambda_0+1+d_0|^2} + \mathrm{Re}\left(\frac{(c_0|\lambda_0|^2+2\overline{\lambda_0})(\overline{\lambda_0}+1+d_0)}{|\lambda_0+1+d_0|^2}\right)$$

$$= \frac{|\lambda_0|^2[1.2(1-c_0)^2+(1+d_0)c_0-2]+1.2(1-d_0)^2}{|\lambda_0+1+d_0|^2}.$$

Thus if $1.2(1-c_0)^2+(1+d_0)c_0-2>0$, we have stability by (5.2). Using $c_0+2d_0=2$, we calculate that this is equivalent to $7c_0^2-4c_0-8>0$, which is true if $c_0>\frac{2}{7}(1+\sqrt{15})\approx 1.3923$.

We compute, using (2.6),

$$c_0 = \frac{2(\cosh\theta_r(L+x_1)-\cosh\theta_r(L-x_1))}{\cosh\theta_r(L+x_1)+\cosh\theta_r(L-x_1)}, \quad d_0 = \frac{2\cosh\theta_r(L-x_1)}{\cosh\theta_r(L+x_1)+\cosh\theta_r(L-x_1)}.$$

Note that for $L=\infty$ we have

$$c_0 = \frac{2(e^{2\theta_r|x_1|}-1)}{e^{2\theta_r|x_1|}+1}, \quad d_0 = \frac{2}{e^{2\theta_r|x_1|}+1}.$$

By (3.17), this implies $\sqrt{\frac{D_s}{D_r}}-1 = e^{2\theta_r|x_1|} > 5.5822$ and $\frac{D_s}{D_r} > 43.33$. If the last condition is valid, we have stability.

We summarize the stability result for the $O(1)$ eigenvalues as follows.

Theorem 5.1. *The mutually exclusive, spiky steady state given in Theorem* 3.2 *is linearly stable with respect to large eigenvalues $\lambda_\epsilon = O(1)$ for $\tau \ge 0$ and $\epsilon>0$ small enough if*

$$\frac{\cosh\theta_r(L+x_1)-\cosh\theta_r(L-x_1)}{\cosh\theta_r(L+x_1)+\cosh\theta_r(L-x_1)} > \frac{1}{7}(1+\sqrt{15}). \tag{5.3}$$

For $L=\infty$, this corresponds to

$$\frac{D_s}{D_r} > 43.33.$$

Now the study of the large eigenvalues is complete. In the next section we study the small eigenvalues.

6. Stability III: The small eigenvalues. Now we study the small eigenvalues for (6.3), namely those with $\lambda_\epsilon \to 0$ as $\epsilon \to 0$. In this section we summarize the main steps and results in several lemmas. Their proofs are rather technical, and we therefore delay them to the appendices.

For given $f \in L^2(\Omega)$, let $T_r[f]$ be the unique solution in $H^2_N(\Omega)$ of the problem

$$D_r \Delta(T_r[f]) - T_r[f] + \alpha_\epsilon f = 0. \tag{6.1}$$

In the same way, the operator T_s is defined with D_r replaced by D_s.

Let

$$\bar{g}_{\epsilon,1} = \hat{t}_1 w_{\epsilon,x_1^\epsilon} + \phi_{\epsilon,x_1^\epsilon}, \quad \bar{g}_{\epsilon,2} = \hat{t}_2 w_{\epsilon,x_2^\epsilon} + \phi_{\epsilon,x_2^\epsilon},$$

$$\bar{r}_\epsilon = cT_r[T_s[\bar{g}_{\epsilon,2}]\bar{g}^2_{\epsilon,1} + T_s[\bar{g}_{\epsilon,1}]\bar{g}^2_{\epsilon,2}], \quad \bar{s}_{\epsilon,1} = T_s[\bar{g}_{\epsilon,1}], \quad \bar{s}_{\epsilon,2} = T_s[\bar{g}_{\epsilon,2}], \tag{6.2}$$

where $\hat{t}_i$ has been defined in (3.7) After rescaling, the eigenvalue problem (4.1) becomes

(6.3)

$$\begin{cases} \lambda_\epsilon \phi_{\epsilon,1} = \epsilon^2 \Delta \phi_{\epsilon,1} - \phi_{\epsilon,1} + \dfrac{c\eta_{\epsilon,2}\bar{g}^2_{\epsilon,1}}{\bar{r}_\epsilon} + \dfrac{2c\bar{s}_{\epsilon,2}\bar{g}_{\epsilon,1}\phi_{\epsilon,1}}{\bar{r}_\epsilon} - \dfrac{c\bar{s}_{\epsilon,2}\bar{g}^2_{\epsilon,1}\psi_\epsilon}{\bar{r}^2_\epsilon}, \\ \lambda_\epsilon \phi_{\epsilon,2} = \epsilon^2 \Delta \phi_{\epsilon,2} - \phi_{\epsilon,2} + \dfrac{c\eta_{\epsilon,1}\bar{g}^2_{\epsilon,2}}{\bar{r}_\epsilon} + \dfrac{2c\bar{s}_{\epsilon,1}\bar{g}_{\epsilon,2}\phi_{\epsilon,2}}{\bar{r}_\epsilon} - \dfrac{c\bar{s}_{\epsilon,1}\bar{g}^2_{\epsilon,2}\psi_\epsilon}{\bar{r}^2_\epsilon}, \\ \tau\lambda_\epsilon \psi_\epsilon = D_r \Delta \psi_\epsilon - \psi_\epsilon + c\alpha_\epsilon \eta_{\epsilon,2}\bar{g}^2_{\epsilon,1} + 2c\alpha_\epsilon \bar{s}_{\epsilon,2}\bar{g}_{\epsilon,1}\phi_{\epsilon,1} + c\alpha_\epsilon \eta_{\epsilon,1}\bar{g}^2_{\epsilon,2} + 2c\alpha_\epsilon \bar{s}_{\epsilon,1}\bar{g}_{\epsilon,2}\phi_{\epsilon,2}, \\ \tau\lambda_\epsilon \eta_{\epsilon,1} = D_s \Delta \eta_{\epsilon,1} - \eta_{\epsilon,1} + \alpha_\epsilon \phi_{\epsilon,1}, \\ \tau\lambda_\epsilon \eta_{\epsilon,2} = D_s \Delta \eta_{\epsilon,2} - \eta_{\epsilon,2} + \alpha_\epsilon \phi_{\epsilon,2}, \end{cases}$$

where all functions are in $H^2_N(\Omega)$ and α_ϵ has been defined in (3.7).

For simplicity, we set $\tau = 0$. Since $\tau\lambda_\epsilon \ll 1$ the results in this section are also valid for τ finite. The case of general $\tau > 0$ can be treated as in [18]. We will see that the small eigenvalues are of the order $O(\epsilon^2)$. To compute them, we will need to expand the eigenfunction up to the order $O(\epsilon)$ term.

Let us define

$$\tilde{g}_{\epsilon,j}(x) = \chi\left(\frac{x - x_j^\epsilon}{r_0}\right)\bar{g}_{\epsilon,j}(x), \quad j = 1, 2, \tag{6.4}$$

where $\chi(x)$ is a smooth cut-off function such that $\chi(x) = 1$ for $|x| < 1$ and $\chi(x) = 0$ for $|x| > 2$. Further,

$$r_0 = \frac{1}{10}\left(1 + x_2, 1 - x_1, \frac{1}{2}|x_1 - x_2|\right). \tag{6.5}$$

In a similar way as in section 3, we define approximate kernel and cokernel, but in contrast now we can use the exact solution given in Theorem 3.2:

$$\mathcal{K}^{new}_{\epsilon,\mathbf{x}^\epsilon} := \text{span}\left\{\epsilon\frac{d}{dx}\tilde{g}_{\epsilon,1}\right\} \oplus \text{span}\left\{\epsilon\frac{d}{dx}\tilde{g}_{\epsilon,2}\right\} \subset (H^2_N(\Omega_\epsilon))^2,$$

$$\mathcal{C}^{new}_{\epsilon,\mathbf{x}^\epsilon} := \text{span}\left\{\epsilon\frac{d}{dx}\tilde{g}_{\epsilon,1}\right\} \oplus \text{span}\left\{\epsilon\frac{d}{dx}\tilde{g}_{\epsilon,2}\right\} \subset (L^2(\Omega_\epsilon))^2,$$

where $\mathbf{x}^\epsilon = (x_1^\epsilon, x_2^\epsilon)$ and $\Omega_\epsilon = \left(-\frac{L}{\epsilon}, \frac{L}{\epsilon}\right)$.

Then it is easy to see that

$$\bar{g}_i(x) = \tilde{g}_{\epsilon,i}(x) + \text{e.s.t.}, \quad i = 1, 2, \tag{6.6}$$

where e.s.t. denotes exponentially small terms.

Note that, by Theorem 3.2, $\tilde{g}_{\epsilon,j}(x) \sim \hat{t}_j w\left(\frac{x-x_j^\epsilon}{\epsilon}\right)$ in $H^2_{loc}(\Omega_\epsilon)$ and $\tilde{g}_{\epsilon,j}$ satisfies

$$\epsilon^2 \Delta \tilde{g}_{\epsilon,j} - \tilde{g}_{\epsilon,j} + \frac{(\tilde{g}_{\epsilon,j})^2 \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} + \text{e.s.t.} = 0, \quad j = 1, 2.$$

Thus $\tilde{g}_{\epsilon,j}^{'} := \frac{d\tilde{g}_{\epsilon,j}}{dx}$ satisfies

$$\epsilon^2 \Delta \tilde{g}_{\epsilon,j}^{'} - \tilde{g}_{\epsilon,j}^{'} + \frac{2c\tilde{g}_{\epsilon,j}\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\tilde{g}_{\epsilon,j}^{'} + \frac{c\tilde{g}_{\epsilon,j}^2}{\bar{r}_\epsilon}\bar{s}_{\epsilon,3-j}^{'} - \frac{c\tilde{g}_{\epsilon,j}^2\bar{s}_{\epsilon,3-j}}{(\bar{r}_\epsilon)^2}\bar{r}_\epsilon^{'} + \text{e.s.t.} = 0. \tag{6.7}$$

Let us now decompose

$$\phi_{\epsilon,j} = \epsilon a_j^\epsilon \tilde{g}_{\epsilon,j}^{'} + \phi_{\epsilon,j}^\perp, \quad j = 1, 2, \tag{6.8}$$

with complex numbers a_j^ϵ, where the factor ϵ is for scaling purposes, to achieve that a_j^ϵ is of order $O(1)$, and

$$\phi_\epsilon^\perp = (\phi_{\epsilon,1}^\perp, \phi_{\epsilon,2}^\perp) \in (\mathcal{K}_{\epsilon,\mathbf{x}^\epsilon}^{new})^\perp,$$

where orthogonality is taken for the scalar product of the product space $(L^2(\Omega_\epsilon))^2$. Note that, by definition,

$$\phi_\epsilon = (\phi_{\epsilon,1}, \phi_{\epsilon,2}) \in \mathcal{K}_{\epsilon,\mathbf{x}^\epsilon}^{new}.$$

Suppose that $\|\phi_\epsilon\|_{H^2(\Omega_\epsilon)} = 1$. Then we need to have $|a_j^\epsilon| \leq C$.

Similarly, we decompose

$$\psi_\epsilon = \epsilon \sum_{j=1}^{2} a_j^\epsilon \psi_{\epsilon,j} + \psi_\epsilon^\perp, \quad \eta_{\epsilon,j} = \epsilon a_j^\epsilon \eta_{\epsilon,j}^0 + \eta_{\epsilon,j}^\perp, \quad j = 1, 2, \tag{6.9}$$

where $\psi_{\epsilon,j}$ satisfies

$$D_r \Delta \psi_{\epsilon,j} - \psi_{\epsilon,j} + 2\alpha_\epsilon c\tilde{g}_{\epsilon,j}\tilde{g}_{\epsilon,j}^{'}\bar{s}_{\epsilon,3-j} + \alpha_\epsilon c\tilde{g}_{\epsilon,3-j}^2 \eta_{\epsilon,j}^0 = 0, \tag{6.10}$$

$\eta_{\epsilon,i}^0$ is given by

$$D_s \Delta \eta_{\epsilon,i}^0 - \eta_{\epsilon,i}^0 + \alpha_\epsilon \tilde{g}_{\epsilon,i}^{'} = 0, \tag{6.11}$$

$\psi_\epsilon^\perp$ satisfies

$$D_r \Delta \psi_\epsilon^\perp - \psi_\epsilon^\perp + 2\alpha_\epsilon\, c\tilde{g}_{\epsilon,1}\bar{s}_{\epsilon,2}\phi_{\epsilon,1}^\perp + \alpha_\epsilon\, c\tilde{g}_{\epsilon,1}^2\eta_{\epsilon,2}^\perp + 2\alpha_\epsilon\, c\tilde{g}_{\epsilon,2}\bar{s}_{\epsilon,1}\phi_{\epsilon,2}^\perp + \alpha_\epsilon\, c\tilde{g}_{\epsilon,2}^2\eta_{\epsilon,1}^\perp = 0, \tag{6.12}$$

and, finally, $\eta_i^\perp$ is given by

$$D_s \Delta \eta_{\epsilon,i}^\perp - \eta_{\epsilon,i}^\perp + \alpha_\epsilon \phi_{\epsilon,i}^\perp = 0. \tag{6.13}$$

Substituting the decompositions of $\phi_{\epsilon,i}$, ψ_ϵ, and $\eta_{\epsilon,i}$ into (6.3), we have

$$\begin{aligned} &\epsilon c\left(a_j^\epsilon \frac{(\tilde{g}_{\epsilon,j})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon^2}\bar{r}_\epsilon^{'} - \sum_{k=1}^{2} a_k^\epsilon \frac{(\tilde{g}_{\epsilon,j})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon^2}\psi_{\epsilon,k}\right) \\ &- \epsilon c\left(a_j^\epsilon \frac{(\tilde{g}_{\epsilon,j})^2}{\bar{r}_\epsilon}\bar{s}_{\epsilon,3-j}^{'} - a_{3-j}^\epsilon \frac{(\tilde{g}_{\epsilon,j})^2}{\bar{r}_\epsilon}\eta_{\epsilon,3-j}^0\right) \\ &+ \epsilon^2 \Delta \phi_{\epsilon,j}^\perp - \phi_{\epsilon,j}^\perp + \frac{2c\tilde{g}_{\epsilon,j}\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\phi_{\epsilon,j}^\perp - \frac{c\tilde{g}_{\epsilon,j}^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon^2}\psi_\epsilon^\perp + \frac{c\tilde{g}_{\epsilon,j}^2}{\bar{r}_\epsilon}\eta_{\epsilon,3-j}^\perp - \lambda_\epsilon\phi_{\epsilon,j}^\perp + \text{e.s.t.} \\ &= \lambda_\epsilon \epsilon a_j^\epsilon \tilde{g}_{\epsilon,j}^{'}, \quad j = 1, 2, \end{aligned} \tag{6.14}$$

since

$$\epsilon^2\Delta\tilde{g}_{\epsilon,j}^{'} - \tilde{g}_{\epsilon,j}^{'} + \frac{2c\tilde{g}_{\epsilon,j}\bar{s}_{3-j,\epsilon}}{\bar{r}_\epsilon}\tilde{g}_{\epsilon,j}^{'} + \text{e.s.t.} = 0.$$

Multiplying both sides of (6.14) for $j = 1, 2$ by $\tilde{g}_{\epsilon,l}^{'}$ for $l = 1, 2$ and integrating over $(-L, L)$, we obtain

$$\text{r.h.s. of (6.14)} = \lambda_\epsilon a_j^\epsilon \epsilon \int_{-L}^{L} \tilde{g}_{\epsilon,j}^{'}\tilde{g}_{\epsilon,l}^{'}\,dx = \lambda_\epsilon \delta_{jl} a_l^\epsilon (\hat{t}_l)^2 \int_R (w^{'}(y))^2\,dy\,(1+o(1)) \tag{6.15}$$

and

$$\begin{aligned}
\text{l.h.s. of (6.14)} &= c\epsilon \sum_{k=1}^{2} a_k^\epsilon \delta_{jl} \int_{-L}^{L} \frac{(\tilde{g}_{\epsilon,j})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon^2}\left(\delta_{jk}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)\tilde{g}_{\epsilon,l}^{'}\,dx \\
&\quad + c\epsilon \sum_{k=1}^{2} a_k^\epsilon \delta_{jl} \int_{-L}^{L} \frac{(\tilde{g}_{\epsilon,j})^2}{\bar{r}_\epsilon}\left(\delta_{j,3-k}\eta_{\epsilon,3-j}^{0} - \delta_{j,k}\bar{s}_{\epsilon,3-j}^{'}\right)\tilde{g}_{\epsilon,l}^{'}\,dx \\
&\quad + c\delta_{jl} \int_{-L}^{L} \frac{(\tilde{g}_{\epsilon,l})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\left(\frac{\bar{r}_\epsilon^{'}}{\bar{r}_\epsilon} - \frac{\bar{s}_{\epsilon,3-j}^{'}}{\bar{s}_{\epsilon,3-j}}\right)\phi_{\epsilon,j}^{\perp}\,dx \\
&\quad + c\delta_{jl} \int_{-L}^{L} \frac{(\tilde{g}_{\epsilon,j})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\left(\frac{\eta_{\epsilon,3-j}^{\perp}}{\bar{s}_{\epsilon,3-j}} - \frac{\psi_\epsilon^{\perp}}{\bar{r}_\epsilon}\right)\tilde{g}_{\epsilon,l}^{'}\,dx + o(\epsilon^2)
\end{aligned}$$

$$= J_{1,l} + J_{2,l} + J_{3,l} + J_{4,l} := J_l, \tag{6.16}$$

where $J_{i,l}$, $i = 1, \ldots, 4$, are defined by the last equality. The following is the key lemma for the asymptotic behavior of the small eigenvalues.

Lemma 6.1. *We have*

$$\begin{aligned}
&J_l \\
&= -\epsilon^2\left(\int_R \frac{1}{3}w^3\,dy\right)\sum_{k=1}^{2} a_k^\epsilon \Bigg\{\Bigg\{-\hat{t}_l\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)\right) + \hat{t}_{3-l}\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon))\Bigg\} \\
&\qquad - \nabla_{x_l^\epsilon}\left(\frac{\delta_{k,3-l}\nabla_{x_{3-l}^\epsilon}G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}\right) \\
&\quad + \Bigg\{(\nabla_{x_k^\epsilon}\hat{t}_l(x_1^\epsilon, x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon, x_l^\epsilon) + (\nabla_{x_k^\epsilon}\hat{t}_{3-l}(x_1^\epsilon, x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\Bigg\}\Bigg\} + o(\epsilon^2).
\end{aligned} \tag{6.17}$$

Lemma 6.1 follows from the following series of lemmas.

Lemma 6.2. *We have*

$$\eta_{\epsilon,k}^{0}(x_{3-k}^\epsilon) = \hat{t}_k\nabla_{x_k^\epsilon}G_{D_s}(x_{3-k}^\epsilon, x_k^\epsilon) + O(\epsilon). \tag{6.18}$$

Lemma 6.3. *We have*

$$\bar{s}_{\epsilon,k}^{'}(x_{3-k}^\epsilon) = \hat{t}_k\nabla_{x_{3-k}^\epsilon}G_{D_s}(x_{3-k}^\epsilon, x_k^\epsilon) + O(\epsilon). \tag{6.19}$$

LEMMA 6.4. *For $k, l = 1, 2$ we have*

$$\begin{aligned}\left(\delta_{kl}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)(x_l^\epsilon) = c\hat{t}_1\hat{t}_2\Big\{ & -\hat{t}_l \nabla_{x_k^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) \\ & + \hat{t}_{3-l}\nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) \\ & + \frac{1}{2\sqrt{D_r}}\hat{t}_l \nabla_{x_k^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\Big\} + O(\epsilon).\end{aligned} \tag{6.20}$$

Similar to Lemma 6.4, we get the next claim.

LEMMA 6.5. *For $k, l = 1, 2$ we have*

$$\begin{aligned}& \left(\delta_{kl}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)(x_l^\epsilon + \epsilon y) - \left(\delta_{kl}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)(x_l^\epsilon) \\ & = \epsilon y c\hat{t}_1\hat{t}_2\Big\{ - \hat{t}_l \nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) \\ & \qquad + \hat{t}_{3-l}\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) \\ & \qquad + \frac{1}{2\sqrt{D_r}}\hat{t}_l \nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\Big\} + O(\epsilon^2).\end{aligned} \tag{6.21}$$

Lemma 6.1 will be shown in Appendix A, proving Lemmas 6.2–6.5 first.

After obtaining the asymptotic behavior of the small eigenvalues, our next goal is to study their stability.

Combining Lemma 6.1 with (6.15) and (6.16), the small eigenvalues λ^ϵ are given by the following two-dimensional eigenvalue problem, where $(a_1^\epsilon, a_2^\epsilon)$ are the corresponding eigenvectors:

$$\begin{aligned}& -\epsilon^2\hat{t}_l\left(\int_R \frac{1}{3}w^3\,dy\right)\sum_{k=1}^2 a_k^\epsilon\Bigg\{\Big\{-\hat{t}_l\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)\right) + \hat{t}_{3-l}\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\Big\} \\ & \qquad\qquad - \nabla_{x_l^\epsilon}\left(\frac{\delta_{k,3-l}\nabla_{x_{3-l}^\epsilon}G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}\right) \\ & + \Big\{(\nabla_{x_k^\epsilon}\hat{t}_l(x_1^\epsilon, x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon, x_l^\epsilon) + (\nabla_{x_k^\epsilon}\hat{t}_{3-l}(x_1^\epsilon, x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\Big\}\Bigg\} + o(\epsilon^2) \\ & \qquad = \lambda_\epsilon\delta_{jl}a_l^\epsilon(\hat{t}_l)^2\int_R (w^{'}(y))^2\,dy\,(1 + o(1)).\end{aligned} \tag{6.22}$$

From (6.22) it follows that the eigenvectors $(a_1^0, a_2^0) = \lim_{\epsilon\to 0}(a_1^\epsilon, a_2^\epsilon)$ satisfy $(a_1^0, a_2^0) = (1, -1)$ or $(a_1^0, a_2^0) = (1, 1)$, up to a constant factor.

For the eigenvector $(a_1^0, a_2^0) = (1, -1)$, the computations of the eigenvalue λ_1^ϵ are similar to those given in section 3. We get

$$\lambda_1^\epsilon = C_3\epsilon^2 M'(x_1^\epsilon) + o(\epsilon^2),$$

where

$$M(x) = -2\theta_s\tanh\theta_s(L - x) + \theta_r\tanh\theta_r(L - x) + \theta_r\frac{\sinh\theta_r(L - x) - \sinh\theta_r(L + x)}{\cosh\theta_r(L - x) + \cosh\theta_r(L + x)}$$

and

$$C_3 = \frac{1}{3\hat{t}_l} \frac{\int_R w^3\, dy}{\int_R (w')^2\, dy} > 0. \tag{6.23}$$

This implies

$$\begin{aligned} M^{'}(x) = {} & \frac{2\theta_s^2}{\cosh^2 \theta_s(L-x)} - \frac{\theta_r^2}{\cosh^2 \theta_r(L-x)} \\ & - \theta_r^2 \left(1 - \frac{[\sinh \theta_r(L-x) - \sinh \theta_r(L+x)]^2}{[\cosh \theta_r(L-x) - \cosh \theta_r(L+x)]^2} \right). \end{aligned}$$

Obviously, $M'(x) < 0$ if $\theta_s = 0$ or if θ_s is small compared to θ_r. A simple sufficient condition is obtained by taking into account the first two terms of $M'(x)$ which has been derived in section 3 and is given by (3.13). Recall that (3.13) holds if $D_s/D_r > 4$.

If $D_s/D_r > 4$, the eigenvalue λ_1^ϵ has negative real part.

Now we consider the eigenvalue λ_2^ϵ with eigenvector such that $\lim_{\epsilon\to 0}(a_1^\epsilon, a_2^\epsilon) = (1,1)$. We have the following result.

Lemma 6.6. *Suppose that* λ_2^ϵ *is the eigenvalue with eigenvector* $\lim_{\epsilon\to 0}(a_1^\epsilon, a_2^\epsilon) = (1,1)$. *Then we have*

$$\lambda_2^\epsilon = C_3 \epsilon^2 P(x_1^\epsilon, x_2^\epsilon) + o(\epsilon^2), \quad \textit{where } C_3 > 0 \textit{ has been defined in } (6.23), \tag{6.24}$$

and

$$\begin{aligned} P(x_1^\epsilon, x_2^\epsilon) = (\nabla_{x_1^\epsilon} + \nabla_{x_2^\epsilon}) \Bigg\{ & \frac{(\nabla_{x_1^\epsilon} - \nabla_{x_2^\epsilon}) G_{D_s}(x_1^\epsilon, x_2^\epsilon)}{G_{D_s}(x_1^\epsilon, x_2^\epsilon)} \\ & - \hat{t}_1^\epsilon(x_1^\epsilon, x_2^\epsilon)(\nabla_{x_1^\epsilon} - \nabla_{x_2^\epsilon}) H_{D_r}(x_1^\epsilon, x_1^\epsilon) - \hat{t}_2^\epsilon(x_1^\epsilon, x_2^\epsilon)(\nabla_{x_1^\epsilon} - \nabla_{x_2^\epsilon}) H_{D_r}(x_1^\epsilon, x_1^\epsilon) \Bigg\}. \end{aligned}$$

We have $P(x_1^\epsilon, x_2^\epsilon) \le 0$ *with equality if and only if* $x_1^\epsilon = x_2^\epsilon = 0$.

Lemma 6.6 will be proved in Appendix B.

By the argument of Dancer [2] the eigenvalue problem (6.22) captures all converging sequences of small eigenvalues λ^ϵ, and so λ_1^ϵ and λ_2^ϵ are all $o(1)$ eigenvalues for ϵ small enough. Therefore we have the following main result on $o(1)$ eigenvalues.

Theorem 6.7. *Suppose* $D_s/D_r > 4$ *and* $\lim_{\epsilon\to 0} x_1^\epsilon = x_1 \neq 0$. *The mutually exclusive, spiky steady state given in Theorem* 3.2 *is linearly stable with respect to small eigenvalues* $\lambda_\epsilon = o(1)$ *if* $\tau \ge 0$ *and* $\epsilon > 0$ *are both small enough. More precisely, we have* $Re(\lambda_\epsilon) \le c\epsilon^2$ *for some* $c > 0$ *independent of* ϵ *and* τ.

7. Numerical simulations. For the simulations we use the domain $\Omega = (-1, 1)$ and Neumann boundary conditions for all components. The constants in the five-component Meinhardt–Gierer system are chosen as follows:

$$\epsilon^2 = .001, \quad D_r = .1, \quad D_s = 1, \quad c = 1, \quad \tau = 1.$$

The graphs in Figure 1 show the numerically obtained long-term limit of the five components g_1, g_2, r, s_1, s_2, i.e., the state at $t = 3{,}000$. After that the solution is numerically stable and does not change anymore. This confirms the analytical result

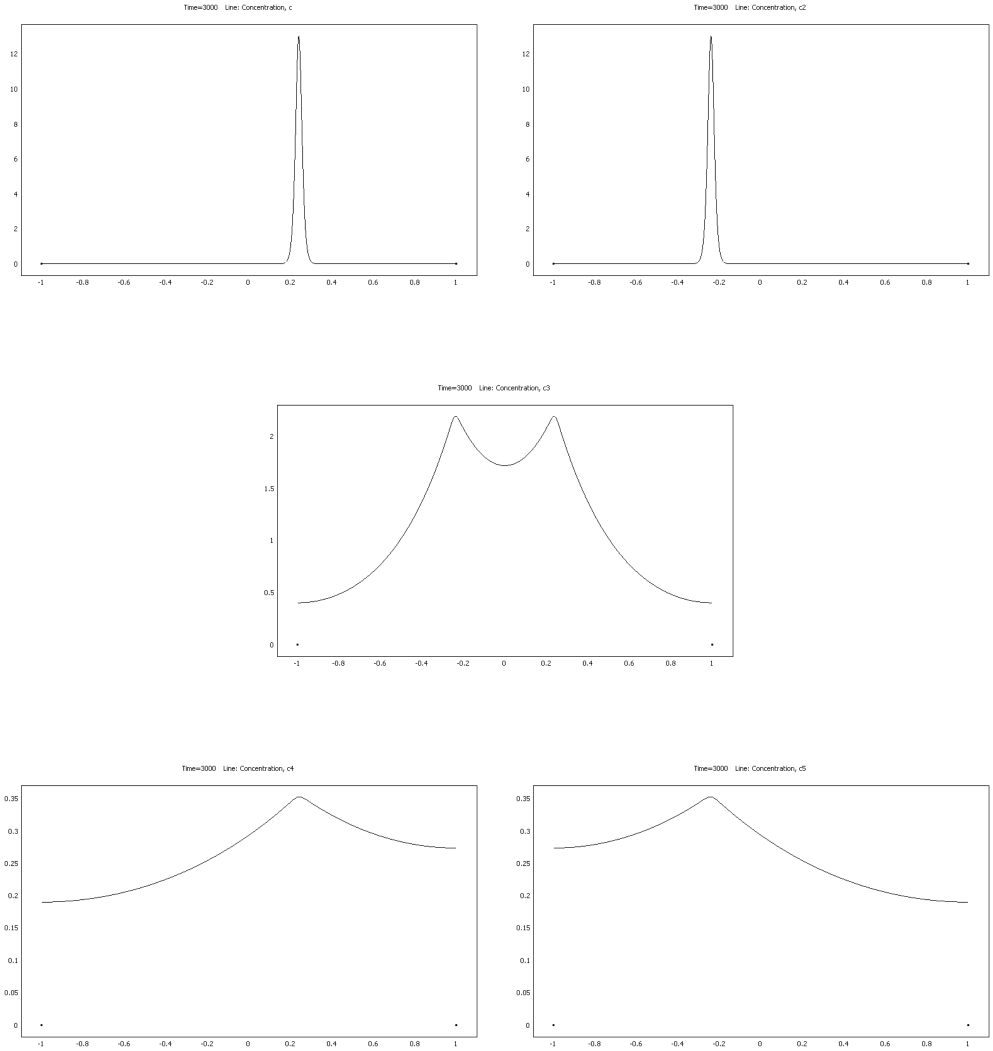


FIG. 1. *The stable, mutually exclusive, two-spike steady state. The five components g_1, g_2, r, s_1, s_2 have been plotted to highlight the interactions between them.*

that the steady state with two mutually exclusive spikes for the two activators which are located in different positions is stable.

Our simulations support the conjecture that the spikes are not only linearly stable as steady states but that, at least locally, they are also dynamically stable for the parabolic reaction-diffusion system.

The choice of constants for the numerical simulations has been motivated by the analysis. In particular, D_r has to be rather small compared to D_s by the stability result in section 4. On the other hand, D_r cannot be too small since otherwise by the results in section 3 the distance between the spikes becomes very large and there is no such solution on the interval $(-1, 1)$. So the parameters have to be chosen very carefully, and without any analytical results it would be very hard to find the parameter range for which stable mutually exclusive spikes exist.

Figure 1 shows that the inhibitor r has two peaks which are near the peaks of the local activators g_1 and g_2. The profile of the peaks of r is "smoother" than for those of the local activators. The lateral activator s_i has a peak near the peak of g_i, and its profile again is smoother than the latter.

We expect Hopf bifurcation and oscillating spikes to occur for sufficiently large tau. We analyzed only the case $\tau = 0$ and did not observe oscillations numerically for $\tau = 1$. The instabilities of the spikes which we encountered in the numerical calculations were (i) disappearance of spikes when their amplitudes becomes unstable (related to large eigenvalues)—this happens if the ratio of the diffusion constants $\frac{D_s}{D_r}$ is too small, and (ii) movement of the spikes to the boundary or towards each other when their positions became unstable (related to small eigenvalues)—this occurs if D_r is too small.

For numerical simulations with very large τ we expect oscillations to occur.

Appendix A. Proof of Lemma 6.1. In this appendix we prove Lemma 6.1 in a sequence of lemmas. First we introduce some notation.

Using the notation of (3.7), we introduce matrix notation

$$e = (1,1)^T, \quad t = \left(\hat{t}_1, \hat{t}_2\right)^T, \quad \nabla_{x_i}\hat{t} = \left(\nabla_{x_i}\hat{t}_1, \nabla_{x_i}\hat{t}_2\right)^T, \quad i = 1,2,$$

$$\mathcal{G}_{ij} = \left(\ G(x_i, x_j)\ \right), \quad i,j = 1,2, \quad \nabla_{x_i}\mathcal{G}_{kl} = \left(\ \nabla_{x_i} G(x_k, x_l)\ \right), \quad i,j,k = 1,2,$$

from which we get

$$\left\{\begin{array}{l} e = \mathcal{G}\hat{t}, \\ 0 = \left(\nabla_{x_1}\mathcal{G}\right)\hat{t} + \mathcal{G}\left(\nabla_{x_1}\hat{t}\right), \\ 0 = \left(\nabla_{x_2}\mathcal{G}\right)\hat{t} + \mathcal{G}\left(\nabla_{x_2}\hat{t}\right). \end{array}\right. \tag{A.1}$$

The system (A.1) has a unique solution $(\hat{t}, \nabla_{x_1}\hat{t}, \nabla_{x_2}\hat{t})$ since $\det(\mathcal{G}) \neq 0$, which can be written as follows:

$$\hat{t} = \mathcal{G}^{-1}e, \quad \nabla_{x_i}\hat{t} = -\mathcal{G}^{-1}\left(\nabla_{x_i}\mathcal{G}\right)\mathcal{G}^{-1}e, \quad i = 1,2. \tag{A.2}$$

Let us set

$$\tilde{L}_{\epsilon,j}\phi_\epsilon^\perp := \epsilon^2\Delta\phi_{\epsilon,j}^\perp - \phi_{\epsilon,j}^\perp + \frac{2c\tilde{g}_{\epsilon,j}\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\phi_{\epsilon,j}^\perp - \frac{c\tilde{g}_{\epsilon,j}^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon^2}\psi_\epsilon^\perp + \frac{c\tilde{g}_{\epsilon,j}^2}{\bar{r}_\epsilon}\eta_{\epsilon,3-j}^\perp \tag{A.3}$$

and $\mathbf{a}_\epsilon := (a_1^\epsilon, a_2^\epsilon)^T$.

We now prove the key lemma, Lemma 6.1, in a sequence of lemmas.

Proof of Lemma 6.2. Note that for $k = 3 - l$ we have

$$\begin{aligned} \eta_{\epsilon,k}^0(x_l^\epsilon) &= \alpha_\epsilon \int_{-L}^{L} G_{D_s}(x_l^\epsilon, z)\tilde{g}_{\epsilon,k}^{'}(z)\,dz + O(\epsilon) \\ &= \alpha_\epsilon\hat{t}_k\nabla_{x_k^\epsilon}G_{D_s}(x_l^\epsilon, x_k^\epsilon)\int_{-L}^{L} zw'\left(\frac{z-x_k}{\epsilon}\right)(z)\,dz \\ &= -\hat{t}_k\nabla_{x_k^\epsilon}G_{D_s}(x_l^\epsilon, x_k^\epsilon)\alpha_\epsilon\left(\epsilon\int_{-\infty}^{\infty} w(y)\,dy\right) + O(\epsilon) \\ &= -\hat{t}_k\nabla_{x_k^\epsilon}G_{D_s}(x_l^\epsilon, x_k^\epsilon) + O(\epsilon). \qquad \square \end{aligned} \tag{A.4}$$

Proof of Lemma 6.3. Note that for $k = 3 - l$ we have

$$
\begin{aligned}
\bar{s}'_{\epsilon,k}(x_l^\epsilon) &= \alpha_\epsilon \nabla_{x_l^\epsilon} \int_{-L}^{L} G_{D_s}(x_l^\epsilon, z)\tilde{g}_{\epsilon,k}(z)\,dz \\
&= \alpha_\epsilon \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_k^\epsilon) \int_{-L}^{L} \hat{t}_k w\left(\frac{z - x_k}{\epsilon}\right)(z)\,dz + O(\epsilon) \\
&= \alpha_\epsilon \hat{t}_k \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_k^\epsilon)\left(\epsilon \int_{-\infty}^{\infty} w(y)\,dy\right) + O(\epsilon) \\
&= \hat{t}_k \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_k^\epsilon) + O(\epsilon). \qquad \Box
\end{aligned}
\tag{A.5}
$$

Proof of Lemma 6.4. We first consider the case $k = l$ and compute $\psi_{\epsilon,l}(x_l^\epsilon)$ as follows:

$$
\begin{aligned}
\psi_{\epsilon,l}(x_l^\epsilon) &= c\alpha_\epsilon \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left(2\tilde{g}'_{\epsilon,l}\tilde{g}_{\epsilon,l}\bar{s}_{\epsilon,3-l} + \tilde{g}^2_{\epsilon,3-l}\eta^0_{\epsilon,l}\right)(z)\,dz + O(\epsilon) \\
&= c(\alpha_\epsilon)^2 \int_{-\infty}^{\infty} K_{D_r}(|z|)\left(2\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\tilde{g}'_{\epsilon,l}(x_l^\epsilon + z)\right)\int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\,dz \\
&\quad - c(\alpha_\epsilon)^2 \int_{-L}^{L} H_{D_r}(x_l^\epsilon, z)\left(\frac{d}{dz}(\tilde{g}_{\epsilon,l}(z))^2\right)\int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\,dz \\
&\quad + c(\alpha_\epsilon)^2 \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left(\tilde{g}_{\epsilon,3-l}(z)\right)^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}'_{\epsilon,l}(y)\,dy\,dz + O(\epsilon) \\
&= c(\alpha_\epsilon)^2 \int_{-\infty}^{\infty} K_{D_r}(|z|)\left(2\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\tilde{g}'_{\epsilon,l}(x_l^\epsilon + z)\right)\int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\,dz \\
&+ \frac{c}{2}\hat{t}_1\hat{t}_2\hat{t}_l\left(\left(\nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon)\right)G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) + c\hat{t}_1\hat{t}_2\hat{t}_l\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)\nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) \\
&\quad - c\hat{t}_1\hat{t}_2\hat{t}_{3-l}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\nabla_{x_l^\epsilon} G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) + O(\epsilon).
\end{aligned}
\tag{A.6}
$$

Next we consider the case $k = 3 - l$ and compute $\psi_{\epsilon,3-l}(x_l^\epsilon)$ as follows:

$$
\begin{aligned}
\psi_{\epsilon,3-l}(x_l^\epsilon) &= c\alpha_\epsilon \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left(2\tilde{g}'_{\epsilon,3-l}\tilde{g}_{\epsilon,3-l}\bar{s}_{\epsilon,l} + \tilde{g}^2_{\epsilon,l}\eta_{\epsilon,3-l}\right)(z)\,dz + O(\epsilon) \\
&= c(\alpha_\epsilon)^2 \int_{-\infty}^{\infty} K_{D_r}(|z|)\left(\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\right)^2 \int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}'_{\epsilon,3-l}(y)\,dy\,dz \\
&\quad + c(\alpha_\epsilon)^2 \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left(\frac{d}{dz}(\tilde{g}_{\epsilon,3-l}(z))^2\right)\int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,l}(y)\,dy\,dz \\
&\quad - c(\alpha_\epsilon)^2 \int_{-L}^{L} H_{D_r}(x_l^\epsilon, z)\left(\tilde{g}_{\epsilon,l}(z)\right)^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}'_{\epsilon,3-l}(y)\,dy\,dz + O(\epsilon) \\
&= c(\alpha_\epsilon)^2 \int_{-\infty}^{\infty} K_{D_r}(|z|)\left(\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\right)^2 \int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}'_{\epsilon,3-l}(y)\,dy\,dz \\
&\qquad + c\hat{t}_1\hat{t}_2\hat{t}_l\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)\nabla_{x_{3-l}^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) \\
&\qquad - c\hat{t}_1\hat{t}_2\hat{t}_{3-l}\left(\left(\nabla_{x_{3-l}^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\right)G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) \\
&\qquad - c\hat{t}_1\hat{t}_2\hat{t}_{3-l}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)\nabla_{x_{3-l}^\epsilon} G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) + O(\epsilon).
\end{aligned}
\tag{A.7}
$$

Next we compute $\bar{r}_\epsilon(x_l^\epsilon)$:

$$\begin{aligned}
\bar{r}_\epsilon(x_l^\epsilon) &= \alpha_\epsilon c \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left(\tilde{g}_{\epsilon,1}^2 \bar{s}_{\epsilon,2} + \tilde{g}_{\epsilon,2}^2 \bar{s}_{\epsilon,1}\right)(z)\,dz + O(\epsilon)\\
&= (\alpha_\epsilon)^2 c \int_{-\infty}^{\infty} K_{D_r}(|z|)\left\{(\tilde{g}_{\epsilon,l}(x_l^\epsilon + z))^2 \int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right\} dz\\
&\quad - (\alpha_\epsilon)^2 c \int_{-L}^{L} H_{D_r}(x_l^\epsilon, z)\left\{(\tilde{g}_{\epsilon,l}(z))^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right\} dz\\
&\quad + (\alpha_\epsilon)^2 c \int_{-L}^{L} G_{D_r}(x_l^\epsilon, z)\left\{(\tilde{g}_{\epsilon,3-l}(z))^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,l}(y)\,dy\right\} dz + O(\epsilon).
\end{aligned}$$

So we have

$$\begin{aligned}
\bar{r}_\epsilon^{'}(x_l^\epsilon) &= (\alpha_\epsilon)^2 c \int_{-\infty}^{\infty} K_{D_r}(|z|)\left\{\left(2\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\tilde{g}_{\epsilon,l}^{'}(x_l^\epsilon + z)\right) \int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right.\\
&\qquad\left. + (\tilde{g}_{\epsilon,l}(x_l^\epsilon + z))^2 \int_{-L}^{L} \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right\} dz\\
&\quad - (\alpha_\epsilon)^2 c \int_{-L}^{L} \nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, z)\,(\tilde{g}_{\epsilon,l}(z))^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\,dz\\
&\quad + (\alpha_\epsilon)^2 c \int_{-L}^{L} \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, z)\,(\tilde{g}_{\epsilon,3-l}(z))^2 \int_{-L}^{L} G_{D_s}(z, y)\tilde{g}_{\epsilon,l}(y)\,dy + O(\epsilon)\\
&= (\alpha_\epsilon)^2 c \int_{-\infty}^{\infty} K_{D_r}(|z|)\left\{\left(2\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\tilde{g}_{\epsilon,l}^{'}(x_l^\epsilon + z)\right) \int_{-L}^{L} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right.\\
&\qquad\left. + (\tilde{g}_{\epsilon,l}(x_l^\epsilon + z))^2 \int_{-L}^{L} \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\right\} dz\\
&\quad - \frac{c}{2}\hat{t}_1\hat{t}_2\hat{t}_l\left((\nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon))G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\\
&\quad + c\hat{t}_1\hat{t}_2\hat{t}_{3-l}\left((\nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon))G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) + O(\epsilon). \qquad \text{(A.8)}
\end{aligned}$$

Now we compute $\left(\delta_{kl}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)(x_l^\epsilon)$. Again we consider the two cases $k = l$ and $k \neq l$ separately.

First, for $k = l$, we get

$$\begin{aligned}
\left(\bar{r}_\epsilon^{'} - \psi_{\epsilon,l}\right)(x_l^\epsilon) &= -c\hat{t}_1\hat{t}_2\hat{t}_l \nabla_{x_l^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\\
&\quad + c\hat{t}_1\hat{t}_2\hat{t}_{3-l} \nabla_{x_l^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right)\\
&\quad + (\alpha_\epsilon)^2 c \int_{-\infty}^{\infty} K_{D_r}(|z|)\,(\tilde{g}_{\epsilon,l}(x_l^\epsilon + z))^2 \int_{-L}^{L} \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\,dy\,dz + O(\epsilon)\\
&= c\hat{t}_1\hat{t}_2\left\{-\hat{t}_l \nabla_{x_l^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon)G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right) + \hat{t}_{3-l} \nabla_{x_l^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon)G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right)\right.\\
&\qquad\left. + \frac{1}{2\sqrt{D_r}}\hat{t}_l \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right\} + O(\epsilon).
\end{aligned}$$

Next we consider the case $k = 3 - l$ and get

$$\begin{aligned}
-\psi_{\epsilon,3-l}(x_l^\epsilon) &= -c\hat{t}_1\hat{t}_2\hat{t}_l \nabla_{x_{3-l}^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\\
&\quad + \hat{t}_1\hat{t}_2\hat{t}_{3-l} \nabla_{x_{3-l}^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right)\\
&\quad + (\alpha_\epsilon)^2 c \int_{-\infty}^{\infty} K_{D_r}(|z|)\left(\tilde{g}_{\epsilon,l}(x_l^\epsilon + z)\right)^2 \int_{-L}^{L} \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon + z, y)\tilde{g}_{\epsilon,3-l}(y)\, dy\, dz + O(\epsilon)\\
&= c\hat{t}_1\hat{t}_2 \Big\{ -\hat{t}_l \nabla_{x_{3-l}^\epsilon}\left(H_{D_r}(x_l^\epsilon, x_l^\epsilon) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\\
&\quad + \hat{t}_{3-l} \nabla_{x_{3-l}^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right)\\
&\quad + \frac{1}{2\sqrt{D_r}} \hat{t}_l \nabla_{x_{3-l}^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} + O(\epsilon).
\end{aligned}$$

This implies (6.20). The proof of Lemma 6.4 is finished. □

Remark. Note that Lemma 6.4 can be written in the simpler way

$$\begin{aligned}
\left(\delta_{kl}\bar{r}_\epsilon^{'} - \psi_{\epsilon,k}\right)(x_l^\epsilon) &= c\hat{t}_1\hat{t}_2 \Big\{ \hat{t}_l \nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_l^\epsilon) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)\right)\\
&\quad + \hat{t}_{3-l} \nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon)\right) \Big\} + O(\epsilon),
\end{aligned} \tag{A.9}$$

with the understanding that at jump discontinuities the derivative is defined as the arithmetic mean of its left-hand and right-hand derivatives.

Proof of Lemma 6.5. The proof of Lemma 6.5 follows along the same lines as that for Lemma 6.4 and is therefore omitted. □

Before we can complete the proof of Lemma 6.1, we first need to study the asymptotic expansion of $\phi_\epsilon^\perp$ as $\epsilon \to 0$. Let us define

$$\begin{aligned}
\phi_\epsilon^1 &= \begin{pmatrix} \phi_{\epsilon,1}^1 \\ \phi_{\epsilon,2}^1 \end{pmatrix}\\
&:= \epsilon a_1^\epsilon \begin{pmatrix} (\nabla_{x_1} t_1) w_1 \\ (\nabla_{x_1} t_2) w_2 \end{pmatrix} + \epsilon a_2^\epsilon \begin{pmatrix} (\nabla_{x_2} t_1) w_1 \\ (\nabla_{x_2} t_2) w_2 \end{pmatrix} + \epsilon \frac{\mathcal{G}^{-1}\mathcal{W}\mathcal{A}_\epsilon^0 \nabla G_{D_s}(x_1, x_2)}{G_{D_s}(x_1, x_2)},
\end{aligned} \tag{A.10}$$

where w_i, $i = 1, 2$, have been defined in (2.3) and

$$\mathcal{A}_\epsilon^0 = \begin{pmatrix} 0 & a_2^\epsilon \\ a_1^\epsilon & 0 \end{pmatrix}, \quad \mathcal{W} = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}.$$

Then we have the following estimate.

Lemma A.1. *For ϵ sufficiently small, it holds that*

$$\|\phi_\epsilon^\perp - \phi_\epsilon^1\|_{(H^2(\Omega_\epsilon))^2} = O(\epsilon^2). \tag{A.11}$$

Proof. To prove Lemma A.1, we first need to derive a relation between $\phi_{\epsilon,j}^\perp$, $\eta_{\epsilon,j}^\perp$, and $\psi_{\epsilon,j}^\perp$. Note that, similarly to the proof of Proposition 3.1 in section 3, it follows that $\tilde{L}_\epsilon$ is uniformly invertible from $(\mathcal{K}^{new}_{\epsilon,\mathbf{x}^\epsilon})^\perp$ to $(\mathcal{C}^{new}_{\epsilon,\mathbf{x}^\epsilon})^\perp$. By this uniform invertibility, we deduce that

$$\|\phi^\perp_\epsilon\|_{(H^2(\Omega_\epsilon))^2} = O(\epsilon), \quad \text{where } \phi^\perp_\epsilon = \left(\phi^\perp_{\epsilon,1}, \phi^\perp_{\epsilon,2}\right)^T \in (\mathcal{K}^{new}_{\epsilon,\mathbf{x}^\epsilon})^\perp. \tag{A.12}$$

Let us cut off and rescale $\phi^\perp_{\epsilon,j}$ as follows: $\tilde{\phi}_{\epsilon,j} = \frac{\phi^\perp_{\epsilon,j}}{\epsilon}\chi\left(\frac{x-x^\epsilon_j}{r_0}\right)$. Then $\phi^\perp_{\epsilon,j} = \epsilon\tilde{\phi}_{\epsilon,j}$ +e.s.t.

Choose $\phi_{\epsilon,j}$ such that $\|\tilde{\phi}_{\epsilon,j}\|_{H^1(R)} = 1$. Then we have, possibly for a subsequence, that $\tilde{\phi}_{\epsilon,j} \to \phi_j$ in $H^1_{loc}(R)$.

By (6.12) and (6.13), $\psi^\perp_\epsilon$ can be represented as follows (the proof is similar to that of Lemma 6.4):

$$\begin{aligned}
\psi^\perp_\epsilon(x^\epsilon_j) &= \epsilon(\alpha_\epsilon)^2 c \sum_{k=1}^{2} \int_{-L}^{L} G_{D_r}(x^\epsilon_j, z)\Bigg\{2\tilde{g}_{\epsilon,k}(z)\tilde{\phi}_{\epsilon,k}(z)\int_{-L}^{L} G_{D_s}(z,y)\tilde{g}_{\epsilon,3-k}(y)\,dy \\
&\qquad + \tilde{g}^2_{\epsilon,k}(z)\int_{-L}^{L} G_{D_s}(z,y)\tilde{\phi}_{\epsilon,3-k}(y)\,dy\Bigg\}\,dz \qquad \text{(A.13)}\\
&= \epsilon\alpha_\epsilon c \sum_{k=1}^{2} G_{D_r}(x^\epsilon_j, x^\epsilon_k)G_{D_s}(x^\epsilon_k, x^\epsilon_{3-k})\left(2\hat{t}_{3-k}\int_{-L}^{L}\tilde{g}_{\epsilon,k}\tilde{\phi}_{\epsilon,k}\,dx + (\hat{t}_k)^2\int_{-L}^{L}\tilde{\phi}_{\epsilon,3-k}\,dx\right) + o(\epsilon) \\
&= \epsilon c \sum_{k=1}^{2} \hat{t}_k G_{D_r}(x^\epsilon_j, x^\epsilon_k)G_{D_s}(x^\epsilon_k, x^\epsilon_{3-k})\left(2\hat{t}_{3-k}\frac{\int_R w\phi_k\,dy}{\int_R w^2\,dy} + \hat{t}_k\frac{\int_R \phi_{3-k}\,dy}{\int_R w\,dy}\right) + o(\epsilon) \\
&= \frac{\epsilon c}{G_{D_r}(x^\epsilon_1, x^\epsilon_1) + G_{D_r}(x^\epsilon_1, x^\epsilon_2)}\Bigg\{G_{D_r}(x^\epsilon_j, x^\epsilon_j)G_{D_s}(x^\epsilon_j, x^\epsilon_{3-j})\left(2\hat{t}_{3-j}\frac{\int_R w\phi_j\,dy}{\int_R w^2\,dy} + \hat{t}_j\frac{\int_R\phi_{3-j}\,dy}{\int_R w\,dy}\right) \\
&\qquad + G_{D_r}(x^\epsilon_j, x^\epsilon_{3-j})G_{D_s}(x^\epsilon_{3-j}, x^\epsilon_j)\left(2\hat{t}_j\frac{\int_R w\phi_{3-j}\,dy}{\int_R w^2\,dy} + \hat{t}_{3-j}\frac{\int_R\phi_j\,dy}{\int_R w\,dy}\right)\Bigg\} + o(\epsilon).
\end{aligned}$$

In the same way, we calculate

$$\begin{aligned}
\eta^\perp_{\epsilon,3-j}(x^\epsilon_j) &= \epsilon\alpha_\epsilon\int_{-L}^{L} G_{D_s}(x^\epsilon_j, z)\tilde{\phi}_{\epsilon,3-j}(z)\,dz \\
&= \epsilon\alpha_\epsilon G_{D_s}(x^\epsilon_j, x^\epsilon_{3-j})\int_{-L}^{L}\tilde{\phi}_{\epsilon,3-j}\,dx + O(\epsilon^2) \\
&= \epsilon G_{D_s}(x^\epsilon_j, x^\epsilon_{3-j})\frac{\int_R\phi_{3-j}\,dy}{\int_R w\,dy} + o(\epsilon) \qquad \text{(A.14)}
\end{aligned}$$

and

$$\eta^\perp_{\epsilon,j}(x^\epsilon_j) = o(\epsilon). \tag{A.15}$$

Substituting (6.18), (6.19), (6.20), (A), and (A.14) into (6.14) and calculating the limit $\epsilon \to 0$ as we have done in section 4, it follows that $\phi = (\phi_1, \phi_2)^T$ satisfies

$$\begin{aligned}
L\phi &= \Delta\phi - \phi + 2w\phi - \left[\mathcal{B}\int\phi + 2\mathcal{C}\left(\int_R w\phi\right)\right]\left(\int_R w\right)^{-1} w^2 \\
&= \hat{t}_1(\mathbf{a}\cdot\nabla\mathcal{G})\mathcal{G}^{-1}ew^2 - \frac{\hat{t}_1\mathcal{A}^0\nabla G_{D_s}(x_1, x_2)}{G_{D_s}(x_1, x_2)}w^2. \qquad \text{(A.16)}
\end{aligned}$$

In the previous calculation we have used (4.9), (4.10), (A.2), the notation

$$\mathbf{a} = (a_1, a_2)^T = \lim_{\epsilon\longrightarrow 0} (a_1^\epsilon, a_2^\epsilon)^T, \quad \mathbf{a}\cdot\nabla = a_1\nabla_{x_1} + a_2\nabla_{x_2}, \quad x_j = \lim_{\epsilon\longrightarrow 0} x_j^\epsilon,\ j = 1, 2,$$

$$\mathcal{A}^0 = \begin{pmatrix} 0 & a_2 \\ a_1 & 0 \end{pmatrix}$$

and (compare section 2)

$$\bar{r}_\epsilon(x_j^\epsilon) = c\hat{t}_1\hat{t}_2 G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon) + O(\epsilon), \quad j = 1, 2, \tag{A.17}$$

$$\bar{s}_{\epsilon,3-j}(x_j^\epsilon) = \hat{t}_{3-j} G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon) + O(\epsilon), \quad j = 1, 2. \tag{A.18}$$

We compute

$$\mathrm{Id} - \mathcal{B} - 2\mathcal{C} = -\frac{1}{G_{D_r}(x_1,x_1) + G_{D_r}(x_1,x_2)} \begin{pmatrix} G_{D_r}(x_1,x_1) & G_{D_r}(x_1,x_2) \\ G_{D_r}(x_1,x_2) & G_{D_r}(x_2,x_2) \end{pmatrix} = -\hat{t}_1\mathcal{G}.$$

By the Fredholm alternative and since $\det(\mathcal{G}) \neq 0$, equation (A.16) has a unique solution ϕ which is given by

$$\phi = -\mathcal{G}^{-1}(\mathbf{a}\cdot\nabla\mathcal{G})\mathcal{G}^{-1}ew + \frac{\mathcal{G}^{-1}\mathcal{A}^0\nabla G_{D_s}(x_1,x_2)}{G_{D_s}(x_1,x_2)}w. \tag{A.19}$$

Now we compare ϕ with ϕ_ϵ^1. By definition and using (A.2), we get

$$\begin{aligned}\phi_\epsilon^1 &= \left(\epsilon\left(a_1^\epsilon\nabla_{x_1^\epsilon}\hat{t}_1 + a_2^\epsilon\nabla_{x_2^\epsilon}\hat{t}_1\right)\tilde{g}_{\epsilon,1},\ \epsilon\left(a_1^\epsilon\nabla_{x_1^\epsilon}\hat{t}_2 + a_2^\epsilon\nabla_{x_2^\epsilon}\hat{t}_2\right)\tilde{g}_{\epsilon,2}\right)^T \\ &\qquad + \epsilon\frac{\mathcal{G}^{-1}\mathcal{W}\mathcal{A}^0\nabla G_{D_s}(x_1,x_2)}{G_{D_s}(x_1,x_2)} \\ &= \epsilon(\mathbf{a}^\epsilon\cdot\nabla_{x^\epsilon}\hat{t})w + \epsilon\frac{\mathcal{G}^{-1}\mathcal{A}^0\nabla G_{D_s}(x_1,x_2)}{G_{D_s}(x_1,x_2)}w + o(\epsilon) \\ &= -\epsilon\mathcal{G}^{-1}(\mathbf{a}\cdot\nabla\mathcal{G})\mathcal{G}^{-1}ew + \epsilon\frac{\mathcal{G}^{-1}\mathcal{A}^0\nabla G_{D_s}(x_1,x_2)}{G_{D_s}(x_1,x_2)}w + o(\epsilon). \qquad \text{(A.20)}\end{aligned}$$

On the other hand, using (A.19) gives

$$\begin{aligned}\phi_\epsilon^\perp &= \epsilon\left(\tilde{\phi}_{\epsilon,1}, \tilde{\phi}_{\epsilon,2}\right)^T + \text{e.s.t.} = \epsilon\left(\phi_j\left(\frac{x - t_j^\epsilon}{\epsilon}\right)\right)_{j=1,2} + o(\epsilon) \\ &= -\epsilon\mathcal{G}^{-1}(\mathbf{a}\cdot\nabla\mathcal{G})\mathcal{G}^{-1}ew + \epsilon\frac{\mathcal{G}^{-1}\mathcal{A}^0\nabla G_{D_s}(x_1,x_2)}{G_{D_s}(x_1,x_2)}w + o(\epsilon). \qquad \text{(A.21)}\end{aligned}$$

From (A.20) and (A.21), it follows that $\phi_\epsilon = \phi_\epsilon^1 + o(1)$. □

Finally, we complete the proof of the key lemma, Lemma 6.1.

Proof of Lemma 6.1. The computation of J_1 follows from Lemmas 6.4 and 6.5 and from (A.17), (A.18). We get

$$J_{1,l} = c\epsilon\sum_{k=1}^{2} a_k^\epsilon\delta_{jl}\int_{-L}^{L}\frac{c(\tilde{g}_{\epsilon,j})^2\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}\left(\delta_{kl}\frac{\bar{r}_\epsilon'}{\bar{r}_\epsilon} - \frac{\psi_{\epsilon,k}}{\bar{r}_\epsilon}\right)\tilde{g}_{\epsilon,l}'\,dx$$

$$= \epsilon \sum_{k=1}^{2} a_k^\epsilon \delta_{jl} \int_{-L}^{L} c(\tilde{g}_{\epsilon,j})^2 \frac{\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}(x_l^\epsilon) \left(\delta_{kl} \frac{\bar{r}_\epsilon^{'}}{\bar{r}_\epsilon} - \frac{\psi_{\epsilon,k}}{\bar{r}_\epsilon} \right) \tilde{g}_{\epsilon,l}^{'} \, dx$$

$$+ \epsilon \sum_{k=1}^{2} a_k^\epsilon \delta_{jl} \int_{-L}^{L} \frac{c(\tilde{g}_{\epsilon,j})^2 \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} \left[\left(\delta_{kl} \frac{\bar{r}_\epsilon^{'}}{\bar{r}_\epsilon} - \frac{\psi_{\epsilon,k}}{\bar{r}_\epsilon} \right) (x_l^\epsilon) \right] \tilde{g}_{\epsilon,l}^{'} \, dx + o(\epsilon^2)$$

$$= -\epsilon^2 \hat{t}_l \left(\int_R \frac{1}{3} w^3 \, dy \right) \sum_{k=1}^{2} a_k^\epsilon \Bigg\{ \nabla_{x_l^\epsilon} \Bigg\{ -\hat{t}_l \nabla_{x_k^\epsilon} \left(H_{D_r}(x_l^\epsilon, x_l^\epsilon) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \right)$$

$$+ \hat{t}_{3-l} \nabla_{x_k^\epsilon} \left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon) \right) + \frac{1}{2\sqrt{D_r}} \hat{t}_l \nabla_{x_k^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\}$$

$$\times \left\{ \left(\hat{t}_l G_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \right) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \right\}^{-1}$$

$$- \Bigg\{ -\hat{t}_l \nabla_{x_k^\epsilon} \left(H_{D_r}(x_l^\epsilon, x_l^\epsilon) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \right)$$

$$+ \hat{t}_{3-l} \nabla_{x_k^\epsilon} \left(G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) G_{D_s}(x_{3-l}^\epsilon, x_l^\epsilon) \right) + \frac{1}{2\sqrt{D_r}} \hat{t}_l \nabla_{x_k^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\}$$

$$\times \Bigg\{ \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\} \Bigg\{ G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\}^{-2} \Bigg\} + o(\epsilon^2)$$

$$= -\epsilon^2 \hat{t}_l \left(\int_R \frac{1}{3} w^3 \, dy \right) \sum_{k=1}^{2} a_k^\epsilon \Bigg\{ \Bigg\{ -\hat{t}_l \nabla_{x_l^\epsilon} \nabla_{x_k^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_l^\epsilon} \nabla_{x_k^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\}$$

$$+ \nabla_{x_l^\epsilon} \left(\frac{\nabla_{x_k^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)} \right)$$

$$- \Bigg\{ -\hat{t}_l \nabla_{x_k^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_k^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\}$$

$$\times \Bigg\{ -\hat{t}_l \nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Bigg\} \Bigg\} + o(\epsilon^2).$$

In the previous computation of $J_{1,l}$ we have used the condition for the positions of the spikes given in the derivation of Theorem 3.2, which implies that $\frac{\bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon}(x_j^\epsilon) = O(\epsilon)$. More precisely, this condition implies that the second line in the previous computation has only a contribution which was included in the error terms. We will use the same condition in the computation of the other $J_{i,l}$ without explicitly mentioning it again.

Similarly, we compute $J_{2,l}$. We get

$$J_{2,l} = \epsilon \sum_{k=1}^{2} a_k^\epsilon \int_{-L}^{L} \frac{c(\tilde{g}_{\epsilon,j})^2 \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} \left(\delta_{3-j,k} \frac{\eta_{\epsilon,3-j}^0}{\bar{s}_{\epsilon,3-j}} - \delta_{jk} \frac{\bar{s}_{\epsilon,3-j}^{'}}{\bar{s}_{\epsilon,3-j}} \right) \tilde{g}_{\epsilon,l}^{'} \, dx$$

$$= -\epsilon \sum_{k=1}^{2} a_k^\epsilon \int_{-L}^{L} \frac{c(\tilde{g}_{\epsilon,j})^2 \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} \left(\frac{(\delta_{jk} \nabla_{x_j^\epsilon} + \delta_{3-j,k} \nabla_{x_{3-j}^\epsilon}) G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon)}{G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon)} \right) \tilde{g}_{\epsilon,l}^{'} \, dx + o(\epsilon^2)$$

$$= \epsilon^2 \hat{t}_l \left(\int_R \frac{1}{3} w^3(y) \, dy \right) \sum_{k=1}^{2} a_k^\epsilon \nabla_{x_l^\epsilon} \left(\frac{(\delta_{kl} \nabla_{x_l^\epsilon} + \delta_{k,3-l} \nabla_{x_{3-l}^\epsilon}) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)} \right) + o(\epsilon^2).$$

Note that we need to have $k = 3 - j$ and $j = l$; otherwise $J_{2,l}$ is of the order $o(\epsilon^2)$.

The estimate $J_{3,l} = o(\epsilon^2)$ follows by the fact that $\phi^{\perp}_{\epsilon,j} \perp \tilde{g}_{\epsilon,j}$.

Next we determine $J_{4,l}$. We compute, using (A), (A.14), and Lemma 7, that

$$
\begin{aligned}
J_{4,l} &= c\delta_{jl} \int_{-L}^{L} \frac{(\tilde{g}_{\epsilon,j})^2 \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} \left(\frac{\eta^{\perp}_{\epsilon,3-j}}{\bar{s}_{\epsilon,3-j}} - \frac{\psi^{\perp}_\epsilon}{\bar{r}_\epsilon} \right) \tilde{g}^{'}_{\epsilon,l}\, dx \\
&= -\epsilon^2 \hat{t}_l \left(\int_R \frac{1}{3} w^3\, dy \right) \sum_{k=1}^{2} a_k^\epsilon \Bigg\{ \Big\{ (\nabla_{x_k^\epsilon} \hat{t}_l(x_1^\epsilon, x_2^\epsilon)) \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_l^\epsilon) \\
&\qquad\qquad + (\nabla_{x_k^\epsilon} \hat{t}_{3-l}(x_1^\epsilon, x_2^\epsilon)) \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \\
&\quad - \Big\{ (\nabla_{x_k^\epsilon} \hat{t}_l(x_1^\epsilon, x_2^\epsilon)) G_{D_r}(x_l^\epsilon, x_l^\epsilon) + (\nabla_{x_k^\epsilon} \hat{t}_{3-l}(x_1^\epsilon, x_2^\epsilon)) G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \\
&\quad \times \Big\{ -\hat{t}_l \nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \Bigg\} + o(\epsilon^2).
\end{aligned}
$$

Here we have used the relation

$$
\int_{-L}^{L} \frac{c\tilde{g}^2_{\epsilon,j} \bar{s}_{\epsilon,3-j}}{\bar{r}_\epsilon} \frac{\eta^{\perp}_{\epsilon,3-j}}{\bar{s}_{\epsilon,3-j}} \epsilon \tilde{g}^{'}_{\epsilon,j}\, dx = o(\epsilon^2),
$$

which follows from the trivial identity

$$
\nabla_{x_l^\epsilon} \left(\frac{G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon)}{G_{D_s}(x_j^\epsilon, x_{3-j}^\epsilon)} \right) = 0.
$$

In a similar way, using the identity

$$
\nabla_{x_l^\epsilon} \left(\frac{\hat{t}_j G_{D_r}(x_j^\epsilon, x_j^\epsilon) + \hat{t}_{3-j} G_{D_r}(x_j^\epsilon, x_{3-j}^\epsilon)}{\hat{t}_j G_{D_r}(x_j^\epsilon, x_j^\epsilon) + \hat{t}_{3-j} G_{D_r}(x_j^\epsilon, x_{3-j}^\epsilon)} \right) = 0,
$$

it can be seen that the contribution of the term $-\epsilon \frac{\mathcal{G}^{-1}\mathcal{W}\mathcal{A}^0 \nabla G_{D_s}(x_1^\epsilon, x_2^\epsilon)}{G_{D_s}(x_1^\epsilon, x_2^\epsilon)}$ in $\psi^{\perp}_\epsilon$ to $J_{4,l}$ is of the order $o(\epsilon^2)$.

Adding $J_{1,l}$, $J_{2,l}$, and $J_{4,l}$, we get

$$
\begin{aligned}
J_l = -\epsilon^2 \hat{t}_l \left(\int_R \frac{1}{3} w^3\, dy \right) \sum_{k=1}^{2} a_k^\epsilon \Bigg\{ &\Big\{ -\hat{t}_l \nabla_{x_l^\epsilon} \nabla_{x_k^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_l^\epsilon} \nabla_{x_k^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \\
&+ \nabla_{x_l^\epsilon} \left(\frac{\delta_{kl} \nabla_{x_l^\epsilon} G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)} \right) \\
&- \Big\{ -\hat{t}_l \nabla_{x_k^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_k^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \\
&\times \Big\{ -\hat{t}_l \nabla_{x_l^\epsilon} H_{D_r}(x_l^\epsilon, x_l^\epsilon) + \hat{t}_{3-l} \nabla_{x_l^\epsilon} G_{D_r}(x_l^\epsilon, x_{3-l}^\epsilon) \Big\} \\
&- \nabla_{x_l^\epsilon} \left(\frac{(\delta_{kl} \nabla_{x_l^\epsilon} + \delta_{k,3-l} \nabla_{x_{3-l}^\epsilon}) G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon, x_{3-l}^\epsilon)} \right)
\end{aligned}
$$

$$+\Big\{\big(\nabla_{x_k^\epsilon}\hat{t}_l(x_1^\epsilon,x_2^\epsilon)\big)\,\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon,x_l^\epsilon)+\big(\nabla_{x_k^\epsilon}\hat{t}_{3-l}(x_1^\epsilon,x_2^\epsilon)\big)\,\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon,x_{3-l}^\epsilon)\Big\}$$

$$-\Big\{\big(\nabla_{x_k^\epsilon}\hat{t}_l(x_1^\epsilon,x_2^\epsilon)\big)\,G_{D_r}(x_l^\epsilon,x_l^\epsilon)+\big(\nabla_{x_k^\epsilon}\hat{t}_{3-l}(x_1^\epsilon,x_2^\epsilon)\big)\,G_{D_r}(x_l^\epsilon,x_{3-l}^\epsilon)\Big\}$$

$$\times\Big\{-\hat{t}_l\nabla_{x_l^\epsilon}H_{D_r}(x_l^\epsilon,x_l^\epsilon)+\hat{t}_{3-l}\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon,x_{3-l}^\epsilon)\Big\}\Bigg\}+o(\epsilon^2).$$

This expression consists of six parts, which are given in lines 1, 2, 3–4, 5, 6, 7–8, respectively. Part 3 is minus part 6 (up to $o(\epsilon^2)$) by (A.1), and they cancel. Part 2 and part 4 cancel partially.

Making these simplifications, we finally get

$$J_l=-\epsilon^2\hat{t}_l\left(\int_R\frac{1}{3}w^3\,dy\right)\sum_{k=1}^{2}a_k^\epsilon\Bigg\{\Big\{-\hat{t}_l\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(H_{D_r}(x_l^\epsilon,x_l^\epsilon)\right)$$

$$+\hat{t}_{3-l}\nabla_{x_l^\epsilon}\nabla_{x_k^\epsilon}\left(G_{D_r}(x_l^\epsilon,x_{3-l}^\epsilon)\right)\Big\}$$

$$-\nabla_{x_l^\epsilon}\left(\frac{\delta_{k,3-l}\nabla_{x_{3-l}^\epsilon}G_{D_s}(x_l^\epsilon,x_{3-l}^\epsilon)}{G_{D_s}(x_l^\epsilon,x_{3-l}^\epsilon)}\right)$$

$$+\Big\{(\nabla_{x_k^\epsilon}\hat{t}_l(x_1^\epsilon,x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon,x_l^\epsilon)+(\nabla_{x_k^\epsilon}\hat{t}_{3-l}(x_1^\epsilon,x_2^\epsilon))\nabla_{x_l^\epsilon}G_{D_r}(x_l^\epsilon,x_{3-l}^\epsilon)\Big\}\Bigg\}+o(\epsilon^2).$$

This finishes the proof of Lemma 6.1. $\square$

Appendix B. Proof of Lemma 6.6.

Proof of Lemma 6.6. We show that

$$P(x_1^\epsilon,x_2^\epsilon)=(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\Bigg\{\frac{(\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})G_{D_s}(x_1^\epsilon,x_2^\epsilon)}{G_{D_s}(x_1^\epsilon,x_2^\epsilon)}$$

$$-\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)(\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})H_{D_r}(x_1^\epsilon,x_1^\epsilon)-\hat{t}_2^\epsilon(x_1^\epsilon,x_2^\epsilon)(\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})H_{D_r}(x_1^\epsilon,x_1^\epsilon)\Bigg\}<0.$$

We compute

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})G_{D_s}(x_1^\epsilon,x_2^\epsilon)=0$$

and

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})(\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})G_{D_s}(x_1^\epsilon,x_2^\epsilon)=((\nabla_{x_1^\epsilon})^2-(\nabla_{x_2^\epsilon})^2)G_{D_s}(x_1^\epsilon,x_2^\epsilon)=0.$$

Therefore, the first term coming from G_{D_s} gives no contribution at all.

Further, we get

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)=\frac{\nabla_{x_2^\epsilon}G_{D_r}(x_2^\epsilon,x_2^\epsilon)}{\det\mathcal{G}}.$$

To simplify the previous expression, we use the identity

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})(\det\mathcal{G})=0, \tag{B.1}$$

which is easy to derive.

Using (B.1), we get

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)=\frac{\nabla_{x_2^\epsilon}G_{D_r}(x_2^\epsilon,x_2^\epsilon)}{\det\mathcal{G}}, \tag{B.2}$$

which gives

$$\begin{aligned}
&-[(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)](\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})G_{D_r}(x_1^\epsilon,x_1^\epsilon)\\
&\quad=-\frac{\nabla_{x_2^\epsilon}G_{D_r}(x_2^\epsilon,x_2^\epsilon)}{\det\mathcal{G}}\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1\epsilon)\\
&\quad=\frac{\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)}{\det\mathcal{G}}\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon).
\end{aligned} \tag{B.3}$$

In analogy to (B.2), we get

$$(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\hat{t}_2^\epsilon(x_1^\epsilon,x_2^\epsilon)=\frac{\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)}{\det\mathcal{G}}, \tag{B.4}$$

which implies

$$-[(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})\hat{t}_2^\epsilon(x_1^\epsilon,x_2^\epsilon)](\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})G_{D_r}(x_1^\epsilon,x_2^\epsilon)=-\frac{\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)}{\det\mathcal{G}}2\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_2^\epsilon). \tag{B.5}$$

Finally, we compute

$$\begin{aligned}
-\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)(\nabla_{x_1^\epsilon}+\nabla_{x_2^\epsilon})(\nabla_{x_1^\epsilon}-\nabla_{x_2^\epsilon})G_{D_r}(x_1^\epsilon,x_1^\epsilon)&=-\hat{t}_1^\epsilon(x_1^\epsilon,x_2^\epsilon)\nabla^2_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)\\
&=-\frac{G_{D_r}(x_2^\epsilon,x_2^\epsilon)-G_{D_r}(x_1^\epsilon,x_2^\epsilon)}{\det\mathcal{G}}\nabla^2_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon).
\end{aligned} \tag{B.6}$$

Now $P(x_1^\epsilon,x_2^\epsilon)$ is given by the sum of (B.3), (B.5), and (B.6).

Using the explicit expression of the Green's function (2.6), we get for the sum of (B.3) and (B.5)

$$\begin{aligned}
&\frac{\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)}{\det\mathcal{G}}\left[\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)-2\nabla_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_2^\epsilon)\right]\\
&\quad=\frac{\theta_r^4}{\sinh^2 2\theta_r L\det\mathcal{G}}\sinh(2\theta_r-x_1^\epsilon)\left[\sinh 2\theta_r x_1^\epsilon+\sinh 2\theta_r(L-x_1^\epsilon)\right].
\end{aligned}$$

For (B.6), we get

$$-\frac{G_{D_r}(x_2^\epsilon,x_2^\epsilon)-G_{D_r}(x_1^\epsilon,x_2^\epsilon)}{\det\mathcal{G}}\nabla^2_{x_1^\epsilon}G_{D_r}(x_1^\epsilon,x_1^\epsilon)$$

$$=-\frac{\theta_r^4}{\sinh^2 2\theta_r L\det\mathcal{G}}\cosh 2\theta_r(L+x_2^\epsilon)\left[\cosh\theta_r(L-x_2^\epsilon)-\cosh\theta_r(L-x_1^\epsilon)\right]2\cosh 2\theta_r x_1^\epsilon.$$

Adding this all up, we get

$$\begin{aligned} P(x_1^\epsilon, x_2^\epsilon) = \frac{\theta_r^4}{\sinh^2 2\theta_r L \det \mathcal{G}} \Bigg\{ & -2\cosh 2\theta_r(L+x_2^\epsilon)\,[\cosh\theta_r(L-x_2^\epsilon) \\ & - \cosh\theta_r(L-x_1^\epsilon)]\cosh 2\theta_r x_1^\epsilon \\ & + \sinh 2\theta_r x_1^\epsilon\,[\sinh 2\theta_r x_1^\epsilon + \sinh 2\theta_r(L-x_1^\epsilon)] \Bigg\} \\ = \frac{\theta_r^4}{\sinh^2 2\theta_r L \det \mathcal{G}} & \Bigg\{ \cosh 2\theta_r L \cdot [1-\cosh 2\theta_r x_1^\epsilon] \Bigg\}. \end{aligned}$$

Note that for $x_1 = \lim_{\epsilon\to 0} x_1^\epsilon$ we have

$$\cosh 2\theta_r L \cdot [1-\cosh 2\theta_r x_1] \leq 0$$

and

$$\cosh 2\theta_r L \cdot [1-\cosh 2\theta_r x_1] = 0 \quad \text{if and only if} \quad x_1 = 0.$$

Therefore, if $x_1 \neq 0$, then for ϵ small enough we have $P(x_1^\epsilon, x_2^\epsilon) < 0$.

This concludes the proof of Lemma 6.6. □

Acknowledgment. M. W. thanks the Department of Mathematics at CUHK for their kind hospitality.

REFERENCES

[1] H. Bohn, *Interkalare Regeneration und segmentale Gradienten bei den Extremitäten von* Leucophaea-*Larven*, Wilhelm Roux Arch., 165 (1970), pp. 303–341.

[2] E. N. Dancer, *On stability and Hopf bifurcations for chemotaxis systems*, Methods Appl. Anal., 8 (2001), pp. 245–256.

[3] M. del Pino, M. Kowalczyk, and X. Chen, *The Gierer-Meinhardt system: The breaking of homoclinics and multi-bump ground states*, Commun. Contemp. Math., 3 (2001), pp. 419–439.

[4] A. Doelman, R. A. Gardner, and T. J. Kaper, *Stability analysis of singular patterns in the* 1*D Gray-Scott model: A matched asymptotics approach*, Phys. D, 122 (1998), pp. 1–36.

[5] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J. Phys., 50 (2001), pp. 443–507.

[6] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik (Berlin), 12 (1972), pp. 30–39.

[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer, Berlin, Heidelberg, New York, 1983.

[8] C. Gui and J. Wei, *Multiple interior peak solutions for some singular perturbation problems*, J. Differential Equations, 158 (1999), pp. 1–27.

[9] D. Iron, M. Ward, and J. Wei, *The stability of spike solutions to the one-dimensional Gierer-Meinhardt model*, Phys. D, 50 (2001), pp. 25–62.

[10] H. Meinhardt, *Models of Biological Pattern Formation*, Academic Press, London, 1982.

[11] H. Meinhardt and A. Gierer, *Generation and regeneration of sequences of structures during morphogenesis*, J. Theoret. Biol., 85 (1980), pp. 429–450.

[12] W. Sun, M. J. Ward, and R. Russell, *The slow dynamics of two-spike solutions for the Gray–Scott and Gierer–Meinhardt systems: Competition and oscillatory instabilities*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 904–953.

[13] A. M. Turing, *The chemical basis of morphogenesis*, Phil. Trans. Roy. Soc. Lond. B, 237 (1952), pp. 37–72.

[14] M. J. Ward and J. Wei, *Hopf bifurcations and oscillatory instabilities of spike solutions for the one-dimensional Gierer-Meinhardt model*, J. Nonlinear Sci., 13 (2003), pp. 209–264.

[15] J. Wei, *On single interior spike solutions of Gierer-Meinhardt system: Uniqueness, spectrum estimates and stability analysis*, European J. Appl. Math., 10 (1999), pp. 353–378.

[16] J. Wei and M. Winter, *Stationary solutions for the Cahn-Hilliard equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 459–492.

[17] J. Wei and M. Winter, *On the two-dimensional Gierer–Meinhardt system with strong coupling*, SIAM J. Math. Anal., 30 (1999), pp. 1241–1263.

[18] J. Wei and M. Winter, *Spikes for the two-dimensional Gierer-Meinhardt system: The weak coupling case*, J. Nonlinear Sci., 11 (2001), pp. 415–458.

[19] J. Wei and M. Winter, *Spikes for the two-dimensional Gierer-Meinhardt system: The strong coupling case*, J. Differential Equations, 178 (2002), pp. 478–518.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 453–472

THERMOPHORESIS DUE TO STRONG TEMPERATURE GRADIENTS[*]

EHUD YARIV[†]

Abstract. We consider a standard thermophoretic configuration, wherein an insulating spherical particle is suspended within a gaseous domain which is bounded between two parallel walls. The walls are maintained at two different temperatures, thereby generating a nonuniform temperature field within the gas. Due to thermal slip, the particle drifts toward the cold wall. Conventional analyses of this problem, starting with the classical work of Epstein [*Z. Physik.*, 54 (1929), pp. 537–563], employ the small-temperature-difference limit. Then, if the particle is small enough, the problem becomes quasi-steady, and the animating effect of the two walls can be represented by a uniform far-field temperature gradient. The corresponding unbounded problem is identical to other slip-generated problems, such as electrophoresis. We focus here upon the general case where the temperature difference is not small. The dependence of the pertinent flow variables upon the absolute temperature prohibits a transformation to a quasi-steady description, whence the transport problem is governed by an unsteady nonlinear process. The small-particle limit is a singular one, wherein the walls cannot be represented by effective far-field conditions. Moreover, the unique structure of the thermal-slip condition implies that inertia and heat-convection effects are of comparable scaling to wall effects. The singular limit is analyzed using inner–outer expansions. In the outer domain, the temperature field is steady to leading-order, but is not described by a uniform gradient. In the inner particle-scale domain, the flow problem is governed by the steady Stokes equations only in the leading order. The transformation between the inner and outer coordinates involves the particle velocity, itself a dependent variable. Using symmetry arguments, we avoid the detailed calculation of the leading-order flow correction, and focus instead upon its effect on the particle thermophoretic velocity. Due to a fortuitous cancellation of terms, Epstein's result remains valid to leading-order analysis. It has been proposed by Kogan, Galkin, and Fridlender [*Sov. Phys. Usp.*, 19 (1976), pp. 420–428] that thermal stresses must be incorporated into all continuum descriptions which apply to flows driven by $O(1)$ temperature differences. Using symmetry arguments, we also analyze the effects of these stresses in the present configuration.

Key words. inner outer expansions, Stokes equations, phoretic motion

AMS subject classifications. 76P05, 76N15

DOI. 10.1137/070711219

1. Introduction.

When solving continuum flows, the common approach is to postulate a no-slip boundary condition over all solid surfaces in contact with the fluid. Theoretical predictions based upon this condition, however, fail to explain two well-known gas experiments which involve nonisothermal solid surfaces. The first experiment was carried out by Tyndall [21], who observed that solid particles drift away from heated surfaces. (This motion is now known as thermophoresis.) Another deviation from the no-slip prediction was observed by Reynolds [19], who exposed closed gas capillaries to temperature gradients and noticed the establishment of a pressure difference between the capillary ends, with higher pressure at the hot end (see also [15, 2]). (This phenomenon was coined "thermal transpiration.")

These experiments have been performed in the continuum regime. According to "conventional" continuum theory, both describe pure heat conduction scenarios,

[*] Received by the editors December 18, 2007; accepted for publication (in revised form) July 1, 2008; published electronically November 14, 2008. This research was supported by the United States—Israel Binational Science Foundation under grant 2004355.

http://www.siam.org/journals/siap/69-2/71121.html

[†] Department of Mathematics, Technion—Israel Institute of Technology, Technion City 32000, Haifa, Israel (yarive@technion.ac.il).

with no mechanism for flow generation (gravity being neglected). Indeed, both the governing balance equations and the boundary conditions (including the ubiquitous no-slip condition) are satisfied by a static solution. In the absence of flow, however, there is no mechanism to push the particle in Tyndall's experiment, and there can be no pressure variation in Reynolds's experiment.

Both phenomena were explained by Maxwell [16] using gas-kinetic theory. Maxwell analyzed the Knudsen layer which lies in at the gas–solid interface, wherein the gas is out of thermodynamic equilibrium, and showed that even if the layer thickness goes to zero, it results in a finite slip velocity $\boldsymbol{v}_S^*$ at its outer edge—where the continuum boundary conditions are prescribed:

$$\boldsymbol{v}_S^* = c_S \frac{\nu^*}{\theta^*} \frac{\partial \theta^*}{\partial \boldsymbol{x}_S^*}. \tag{1.1}$$

Here c_S is an $O(1)$ dimensionless coefficient, ν^* is the kinematic viscosity, θ^* is the temperature, and $\partial/\partial \boldsymbol{x}_S^*$ denotes surface gradient (dimensional variables are decorated with stars). Maxwell obtained the value $c_S = 3/4$ using the artificial-yet-simple model of central molecular interactions that decay with the sixth power of the intermolecular separation ("Maxwell molecules"). Other systematic analyses of this problem [20, 13] yield different values of c_S.[1] The predicted value of c_S depends upon the molecular model which is employed to solve Boltzmann's equation in the Knudsen layer. At present, there is no universally accepted value for c_S. Practically speaking, however, all molecular models predict $O(1)$ slip coefficients, in agreement with experimental results.

Given the relative slip of gases along nonisothermal solid surfaces, it becomes clear that a force-free solid particle will drift toward cold regions (thermophoresis). In the case of nonuniformly heated closed capillary, the thermal slip along the capillary walls tends to generate a uniform plug flow, directed from the cold end to the hot one. This flow generates a nonzero net mass flux; in a closed capillary, it must therefore be accompanied by a counter-balancing Poiseuillian velocity profile (which is the only nontrivial solution that satisfies the Navier–Stokes equations and a no-slip condition on the capillary walls). This profile, in turn, generates a longitudinal pressure gradient, which is directed toward the hot end of the capillary. This pressure gradient explains the thermal transpiration phenomenon.

In analyzing slip-driven flows, it is a common practice to assume that the temperature difference $\Delta\theta^*$ associated with the imposed wall- temperature distribution is small compared with a characteristic temperature, say θ_∞^*, of that distribution. Then, the gas density and transport coefficients are considered constant. This leads to two major simplifications: the temperature field is essentially governed by conduction (and therefore satisfies Laplace's equation), and the slip condition (1.1) adopts the approximated form

$$\boldsymbol{v}_S^* \approx c_S \frac{\nu_\infty^*}{\theta_\infty^*} \frac{\partial \theta^*}{\partial \boldsymbol{x}_S^*}$$

(where ν_∞^* is a characteristic viscosity value). Within this approximation, the thermal-slip condition is analogous to that appearing in a variety of other phoretic mecha-

[1] The structure (1.1) is valid for arbitrary temperature gradients [14]. The above-mentioned analyses were performed using a linearized version of the Boltzmann equation, valid for small gradients: Since the value of c_S is independent of the magnitude of the gradients, this linearized approach is legitimate.

nisms [1], where the slip velocity is proportional to the gradient of some harmonic field (electric potential, solute concentration, etc.).

The analogy breaks down when $\Delta\theta^*$ is comparable to θ^*_∞. Indeed, numerical simulations in the strong-gradient regime [18] exhibit rich phenomena, not all of which can be described by the linearized model. Despite its importance for a variety of applications (in Annis's experiments [2], for example, $\Delta\theta^* = 300\,\mathrm{K}$), this general case has not yet been systematically investigated. It is the goal of this paper to present an analytic investigation of flows driven by significant temperature differences, where the parameter

$$\zeta = \frac{\Delta\theta^*}{\theta^*_\infty} \tag{1.2}$$

is $O(1)$.

A fundamental property of such thermally-driven flows, identified by Kogan [13], has to do with the velocity scaling. According to the slip condition (1.1), slip-driven flows are quantified by the velocity scale

$$\mathscr{U} = \zeta \frac{\nu^*_\infty}{L}. \tag{1.3}$$

Here, ν^*_∞ is a characteristic viscosity and L is the length dimension associated with the temperature gradient. This scaling implies that the Reynolds number, $\mathscr{U} L/\nu^*_\infty$, is simply given by ζ. Thus, flows which are driven by strong temperature differences are characterized by inherently $O(1)$ Reynolds numbers. (Since the Prandtl number of a gas is $O(1)$, the same conclusion holds for the Péclet number.) This universal scaling contrasts with phoretic slip-driven mechanisms which are characterized by independent velocity scales.[2]

It becomes evident that flows driven by strong temperature differences are inherently nonlinear. The nonlinearity results from both the dependence of thermodynamic and transport properties upon the temperature and from the universal velocity scaling. In a previous work [23], the role of nonlinearity was investigated in the context of channel flows. In this paper we investigate its implications upon thermophoretic particle motion.

Conceptually, the simplest thermophoretic configuration consists of a gaseous domain which is bounded between two walls. The walls are separated distance $2L$ apart, and are held at two different temperatures, say $\theta^*_\infty \pm \Delta\theta^*$. A particle of dimension a is introduced into the gas domain, and drifts toward the cold wall. Practical interest often surrounds the small-particle case, $\epsilon = a/L \ll 1$.

This problem was solved by Epstein [8] in the limit $\zeta \to 0$. In that limit, the flow and heat transport are affected only by the temperature gradients, rather than the absolute temperature. In that case, it is possible to transform the bounded-fluid-domain problem into a comparable unbounded-fluid-domain problem, whereby the two walls are represented by a far-field "imposed gradient" condition. The resulting steady problem is identical to that describing electrophoresis of a constant zeta-potential particle under the action of a uniformly applied electric field [1], the temperature being analogous to the electric potential.

For $\zeta = O(1)$, however, the absolute temperature affects the flow and heat-transport processes. Accordingly, the spatial distribution of the pertinent fields depends upon the instantaneous position of the particle relative to the walls. This

[2] These reflect the dependence of phoretic slip upon dimensional physicochemical surface properties (e.g., surface charge density), which have no counterpart in the thermal-slip mechanism.

dependence implies that the $\zeta = O(1)$ regime is inherently unsteady. Moreover, even for small particles ($\epsilon \ll 1$) it is not possible to replace the two walls by a far-field condition. The limit $\epsilon \ll 1$ is a singular one.

The asymptotic limit $\epsilon \ll 1$ is addressed here via the use of matched asymptotic expansions. The transport processes are analyzed in two separate regions. An "outer" region, characterized by the large length scale L associated with the interwall separation, and an "inner" region, characterized by the particle dimension a. In both regions, each of the pertinent fields is expanded into an asymptotic series in terms of the aspect-ratio parameter ϵ.

The outer region is dominated by a pure heat conduction process. In a sense, this region constitutes the analog of the "far-field" in the linearized problem. In the present problem, however, the outer temperature field is affected by the gas conductivity and therefore is not harmonic; specifically, its longitudinal variation is nonlinear. A spatially-linear profile is attained only in the limit $\zeta \ll 1$, and only then does the outer solution conform to the "applied-gradient" notion.

The inner region is suitable for the description of the transport processes at the particle scale, which are governed by nonlinear and unsteady equations. In that scale, where the fluid domain appears unbounded, conditions at "infinity" are supplemented by the requirement of matching with the outer solution. The equations are closed by imposing Newton's second law on the particle motion, thereby relating the particle velocity to the flow field about it. In view of that dependence, the particle velocity must also be expressed as an asymptotic series in ϵ. Since the transformation between the inner and outer coordinates is performed using this velocity, Van Dyke's matching rules cannot be applied and we resort to the use of intermediate variables.

The leading-order inner flow problem is similar to the linearized $\zeta \ll 1$ problem analyzed by Epstein [8]. Due to the time-varying "applied gradient" appearing in the $\zeta = O(1)$ case, the similarity is not complete. Nevertheless, a fortuitous cancellation of terms reveals that the leading-order thermophoretic velocity is identical to that obtained by Epstein. Nonlinear effects in the flow problem appear only in the leading-order asymptotic correction. This correction is governed by a perturbation of the compressible Navier–Stokes equations. The mathematical problem governing that correction is transformed into a nonhomogeneous creeping-flow equation. Without solving the flow correction in detail, we employ symmetry arguments and show that it does not affect the particle velocity.

We also consider the role of Burnett stresses. In a formal small-Knudsen-number analysis of the Boltzmann equation [6], these stresses constitute a correction to the "conventional" Newtonian stress. Given their small magnitude in continuum flows, Burnett stresses can be ignored in most practical scenarios. Continuum flow in the presence of strong temperature gradients, however, may pose an exceptional case, since two of the Burnett terms (the "thermal stresses") are associated with temperature variations. Indeed, it was shown by Kogan and his coworkers [13, 14] that thermal stresses can actually *generate* flow if such variations are externally imposed. In that case, Burnett terms actually possess the same scaling as the Newtonian stress, and must therefore be superimposed upon the conventional Navier–Stokes equations. A systematic discussion of this asymptotic reordering appears in [4].

Flows driven by strong temperature variations through the action of thermal stresses were studied for a variety of idealized configurations [14], which are usually characterized by several isothermal surfaces (held at different temperatures). In such configurations, the only mechanism for flow generation are the thermal stresses, since no thermal slip is generated on the surfaces. Unfortunately, these idealized configurations are not encountered in practical devices. To this day, as a matter of fact, there is no available experimental evidence for the existence of thermal stresses.

Since the velocity scaling associated with thermal-stress-driven flows is the same as that of flows driven by thermal slip, thermal stresses may be pertinent to the present problem. We therefore revisit the thermophoretic analysis, incorporating the two Burnett thermal terms into the momentum balance. Symmetry arguments show that this incorporation does not affect the $O(\epsilon)$ correction to the particle velocity.

2. Problem formulation. Consider the simplest model of thermophoresis, taking place within an ideal gas domain (constant specific heats c_P and c_V) which is bounded between two parallel solid walls separated distance $2L$ apart. The two walls are maintained at two different temperatures, say $(1 \pm \zeta)\,\theta^*_\infty$ (where $0 < \zeta < 1$).

Neglecting gravity effects, the pressure field between the walls must be uniform; it is denoted by p^*_∞. Since the temperature field θ^* is nonuniform, so must be the density field ρ^*. The static pure-conduction state described here is compatible with the equations of fluid motion.

A spherical solid particle of radius a ($a < L$) is now introduced between the walls. For simplicity, we assume a thermally-insulating particle. The nonuniform temperature field along the particle surface, in conjunction with the slip condition (1.1), implies that the particle-free static state is perturbed: the slip animates a velocity field $\boldsymbol{v}^*$, and a consequent modification of the the uniform pressure p^*_∞ to a nonuniform distribution p^*; the fields θ^* and ρ^* are then modified from their respective static distributions. Since the particle is freely suspended, these fields may result in its motion relative to the ambient nonuniform temperature field. The problem is therefore inherently unsteady; in general, then, all fields depend upon both the position vector $\boldsymbol{x}^*$ and the time t^*.

Clearly, the problem is axi-symmetric about an axis (say z^*) which runs perpendicular to the walls and passes through the particle center. For convenience, we take the walls to be at $z^* = \pm L$. Symmetry implies that the particle velocity is given by $\boldsymbol{w}^* = \boldsymbol{e}_z\, w^*$, $\boldsymbol{e}_z$ being a unit vector pointing in the positive-z^* direction. The instantaneous configuration of the system is determined by the position $z^*_P\,(t^*)$ of the particle center, where $w^* = dz^*_P/dt^*$. A schematic of the problem is presented in Figure 2.1.

The pertinent fields ($\boldsymbol{v}^*$, p^*, ρ^*, and θ^*) involved in this unsteady transport

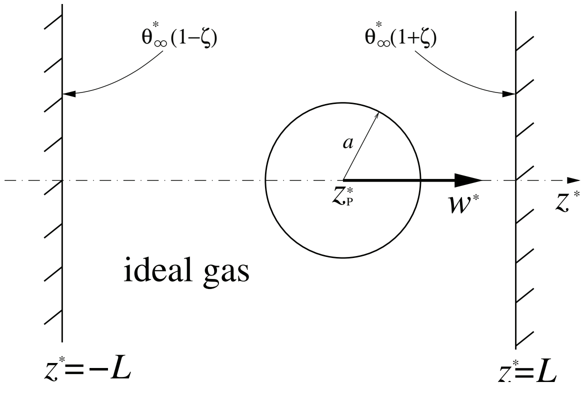


FIG. 2.1. *A schematic of the particle-walls configuration.*

process are governed by the standard set of balance equations [3], in which the viscosity μ^* and the heat conductivity k^* are both temperature dependent. The heat transport is described by the enthalpy-balance equation,

$$c_P \rho^* \frac{D\theta^*}{Dt^*} = \frac{\partial}{\partial \boldsymbol{x}^*} \cdot \left(k^* \frac{\partial \theta^*}{\partial \boldsymbol{x}^*} \right) + \frac{Dp^*}{Dt^*} + \Phi^*. \tag{2.1}$$

Here,

$$\frac{D}{Dt^*} = \frac{\partial}{\partial t^*} + \boldsymbol{v}^* \cdot \frac{\partial}{\partial \boldsymbol{x}^*}$$

is the material derivative operator, and

$$\Phi^* = 2\mu^* \mathsf{e}^* : \mathsf{e}^* \tag{2.2}$$

is the dissipation rate, wherein e^* is the rate-of-strain tensor:

$$\mathsf{e}^* = \frac{1}{2} \left[\left(\frac{\partial \boldsymbol{v}^*}{\partial \boldsymbol{x}^*} \right) + \left(\frac{\partial \boldsymbol{v}^*}{\partial \boldsymbol{x}^*} \right)^\dagger \right] - \frac{1}{3} \left(\frac{\partial}{\partial \boldsymbol{x}^*} \cdot \boldsymbol{v}^* \right) \mathsf{I}. \tag{2.3}$$

The flow is described by the mass-balance equation

$$\frac{D\rho^*}{Dt^*} + \rho^* \frac{\partial}{\partial \boldsymbol{x}^*} \cdot \boldsymbol{v}^* = 0 \tag{2.4}$$

and the momentum-balance equation

$$\rho^* \frac{D\boldsymbol{v}^*}{Dt^*} = -\frac{\partial p^*}{\partial \boldsymbol{x}^*} + 2 \frac{\partial}{\partial \boldsymbol{x}^*} \cdot (\mu^* \mathsf{e}^*). \tag{2.5}$$

The three thermodynamic fields are coupled by the ideal-gas equation of state,

$$p^* = \mathscr{R} \rho^* \theta^*. \tag{2.6}$$

Here, $\mathscr{R} = c_P - c_V$ is the gas constant. Also, the dependence of μ^* and k^* upon θ^* is provided by the functional relations:

$$\frac{\mu^*}{\mu^*_\infty} = f_\mu \left(\frac{\theta^*}{\theta^*_\infty} \right), \quad \frac{k^*}{k^*_\infty} = f_k \left(\frac{\theta^*}{\theta^*_\infty} \right), \tag{2.7}$$

where μ^*_∞ and k^*_∞ are the values of μ^* and k^* at the reference temperature θ^*_∞ (so that $f_\mu(1) = f_k(1) = 1$).

The differential equations are supplemented by the appropriate boundary conditions. On the two walls $z^* = \pm L$ the imposed thermal conditions are

$$\theta^* = \theta^*_\infty (1 \pm \zeta), \tag{2.8}$$

and the no-slip condition is

$$\boldsymbol{v}^* = \boldsymbol{0}. \tag{2.9}$$

The conditions on the particle surface are written using the relative position vector,

$$\boldsymbol{r}^* = \boldsymbol{x}^* - \boldsymbol{e}_z \, z^*_P \, (t^*) \, , \tag{2.10}$$

which is measured from the particle center. Thus, on $r^* = a$ the temperature field satisfies the no-flux condition,

$$\frac{\partial \theta^*}{\partial n} = 0, \tag{2.11}$$

and the velocity field satisfies the slip condition (cf. (1.1))[3]

$$\boldsymbol{v}^* - \boldsymbol{w}^* = c_S \frac{\nu^*}{\theta^*} \frac{\partial \theta^*}{\partial \boldsymbol{x}^*}. \tag{2.12}$$

At large distances from the particle, the flow-induced pressure disturbance attenuates, and the pressure approaches its quiescent value:

$$p^* \to p^*_\infty \quad \text{as} \quad {x^*}^2 + {y^*}^2 \to \infty. \tag{2.13}$$

The governing equations are closed by imposing Newton's second law on the particle motion:

$$\text{hydrodynamic force on particle} = \frac{4\pi}{3} \sigma \rho^*_\infty a^3 \frac{dw^*}{dt^*}. \tag{2.14}$$

Here, $\rho^*_\infty = p^*_\infty / \mathscr{R} \theta^*_\infty$ is a reference density value, and $\sigma \rho^*_\infty$ is the average particle density.

3. Dimensionless formulation. In what follows, it proves convenient to employ dimensionless variables, which appear without the star designation. The coordinates and gradient operator are normalized using the length L. The normalized density and pressure fields are given by the ratios

$$p = \frac{p^*}{p^*_\infty}, \quad \rho = \frac{\rho^*}{\rho^*_\infty}, \tag{3.1}$$

and the reduced temperature θ is defined by the relation

$$\frac{\theta^*}{\theta^*_\infty} = 1 + \zeta \theta. \tag{3.2}$$

Velocity variables are normalized using the velocity scale (see (1.3)):

$$\mathscr{U} = \zeta \frac{\mu^*_\infty}{\rho^*_\infty L}. \tag{3.3}$$

In the transition to dimensionless description, three parameters emerge: (i) the Prandtl number $\mathrm{Pr} = \mu^*_\infty c_P / k^*_\infty$; (ii) the ratio $\gamma = c_P / c_V$; and (iii) the Mach number $M = \mathscr{U} / c^*_\infty$, wherein c_∞ is the sound speed in the reference state:

$$c^2_\infty = \gamma \mathscr{R} \theta^*_\infty. \tag{3.4}$$

The dimensionless enthalpy equation adopts the form

$$\zeta \rho \frac{D\theta}{Dt} = \frac{2}{\mathrm{Pr}} \frac{\partial}{\partial \boldsymbol{x}} \cdot \left[f_k \left(1 + \zeta\theta\right) \frac{\partial \theta}{\partial \boldsymbol{x}} \right] + \frac{\gamma - 1}{\gamma} \frac{Dp}{Dt} + \frac{\gamma - 1}{\zeta} M^2 \Phi, \tag{3.5}$$

[3] In principle, (2.12) should be written using a surface gradient operator; (2.11), however, implies that the conventional gradient operator is equivalent here.

with $D/Dt = \partial/\partial t + \boldsymbol{v}\cdot\partial/\partial\boldsymbol{x}$. Here, the dimensionless dissipation is

$$\Phi = 2f_\mu\,(1+\zeta\theta)\,\mathsf{e}:\mathsf{e}, \tag{3.6}$$

where e is given by the dimensionless equivalent of (2.3). The dimensionless mass- and momentum-conservation equations are

$$\frac{D\rho}{Dt} + \rho\frac{\partial}{\partial\boldsymbol{x}}\cdot\boldsymbol{v} = 0 \tag{3.7}$$

and

$$\zeta\rho\frac{D\boldsymbol{v}}{Dt} = -\frac{\zeta}{\gamma}M^{-2}\frac{\partial p}{\partial\boldsymbol{x}} + 2\frac{\partial}{\partial\boldsymbol{x}}\cdot\left[f_\mu\,(1+\zeta\theta)\,\mathsf{e}\right]. \tag{3.8}$$

Last, the dimensionless equation of state appears as

$$p = \rho\,(1+\zeta\theta)\,. \tag{3.9}$$

Given the velocity scaling of thermal slip, it is not surprising that both the Reynolds and Péclet numbers are simply given by ζ.

The boundary conditions on the two walls are

$$\left.\begin{array}{c}\theta = \pm 1,\\ \boldsymbol{v} = \boldsymbol{0},\end{array}\right\} \quad \text{at} \quad z = \pm 1. \tag{3.10}$$

At the particle surface, the no-flux condition reads as

$$\frac{\partial\theta}{\partial r} = 0 \quad \text{at} \quad r = \epsilon \tag{3.11}$$

and the slip condition is

$$\boldsymbol{v} - \boldsymbol{w} = \frac{c_S}{\rho}\frac{f_\mu\,(1+\zeta\theta)}{1+\zeta\theta}\frac{\partial\theta}{\partial\boldsymbol{x}} \quad \text{at} \quad r = \epsilon. \tag{3.12}$$

At large distances from the particle, the pressure approaches the unperturbed value

$$p \to 1 \quad \text{as} \quad x^2 + y^2 \to \infty. \tag{3.13}$$

The equations are closed by the dimensionless version of (2.14).

4. The dynamic incompressibility limit. In continuum gas flows, the Knudsen number Kn must be small [6]. This number is proportional to the ratio of the Mach number to the Reynolds number. In the present context, the latter is given by the scaled temperature difference ζ, which is $O(1)$. Accordingly, the Mach number is small.

We therefore extract the leading-order limit of the preceding equations for $M \ll 1$. In that limit, the momentum-balance (3.8) and the far-field condition (3.13) imply the following distinguished limit for the pressure field:

$$p \to 1 + \frac{\gamma}{\zeta}M^2\tilde{p}. \tag{4.1}$$

The "dynamic pressure" $\tilde{p}$ actually represents a Stokes-type normalization using the viscous scale $\mu^*_\infty\mathscr{U}/L$ (cf. the "thermodynamic" normalization (3.1)):

$$\tilde{p} = \frac{p^* - p^*_\infty}{\mu^*_\infty\mathscr{U}/L}. \tag{4.2}$$

Since the pressure is constant to leading-order (the "hydrodynamic" pressure is unaffected by the flow), the equation-of-state (3.9) becomes

$$\rho = \frac{1}{1+\zeta\theta}. \tag{4.3}$$

This equation describes a "dynamically-incompressible" fluid [3], whose density is affected only by the temperature.

The density is therefore eliminated as an independent variable. Thus, the slip condition (3.12) becomes

$$\boldsymbol{v} - \boldsymbol{w} = c_S f_\mu \left(1+\zeta\theta\right) \frac{\partial\theta}{\partial\boldsymbol{x}} \quad \text{at} \quad r = \epsilon, \tag{4.4}$$

and the balance equations (3.5), (3.7), and (3.8) adopt the following continuum limits:

$$\frac{\zeta \,\mathrm{Pr}}{1+\zeta\theta}\frac{D\theta}{Dt} = \frac{\partial}{\partial\boldsymbol{x}} \cdot \left[f_k \left(1+\zeta\theta\right) \frac{\partial\theta}{\partial\boldsymbol{x}}\right], \tag{4.5}$$

$$\frac{\partial}{\partial\boldsymbol{x}} \cdot \boldsymbol{v} = \frac{\zeta}{1+\zeta\theta}\frac{D\theta}{Dt}, \tag{4.6}$$

$$\frac{\zeta}{1+\zeta\theta}\frac{D\boldsymbol{v}}{Dt} = -\frac{\partial\tilde{p}}{\partial\boldsymbol{x}} + 2\frac{\partial}{\partial\boldsymbol{x}} \cdot \left[f_\mu \left(1+\zeta\theta\right) \mathsf{e}\right]. \tag{4.7}$$

Note that both the pressure and dissipation terms disappear from the enthalpy equation, which adopts a standard convective–diffusive form. Also, because the fluid is dynamically incompressible, it is only the gradient of the dynamic pressure $\tilde{p}$ which affects the flow. Since the hydrodynamic force is insensitive to the addition of a constant pressure, the pressure $\tilde{p}$ is effectively defined up to an additive constant.

The preceding analysis resembles the Janzen–Rayleigh expansions for small Mach number inviscid flows. Similar analyses for thermodynamically-compressible flows in the limit of small Mach numbers were also performed in the context of forced convection [5, 10].

The present framework of strong-gradient thermophoresis introduces two new features which are absent in the linearized model: (i) the Péclet and Reynolds numbers are both $O\left(1\right)$—in principle, nonlinear convective terms are of the same order as the diffusive terms; (ii) since the absolute temperature affects the transport process, the flow depends upon the instantaneous location of the particle, and the problem is inherently unsteady.

For simplicity of subsequent analysis, we assume Maxwellian intermolecular interactions. Thus, $\mathrm{Pr} = 2/3$, and the transport coefficients are proportional to the temperature

$$f_\mu\left(\eta\right) = f_k\left(\eta\right) = \eta. \tag{4.8}$$

5. Small-particle limit. We now focus upon the small-particle case, $\epsilon = a/L \ll 1$. Since the flow is driven by a slip mechanism at the particle scale, the actual Reynolds and Péclet numbers that characterize the transport are modified from $O(1)$ to $O\left(\epsilon\right)$, whence the continuum domain is modified from $M \ll 1$ to $M \ll \epsilon$. Thus, it is still consistent to employ the preceding equations of a dynamically incompressible gas, obtained via neglecting $O\left(M^2\right)$ terms, while retaining both $O\left(\epsilon\right)$ and $O\left(\epsilon^2\right)$ corrections.

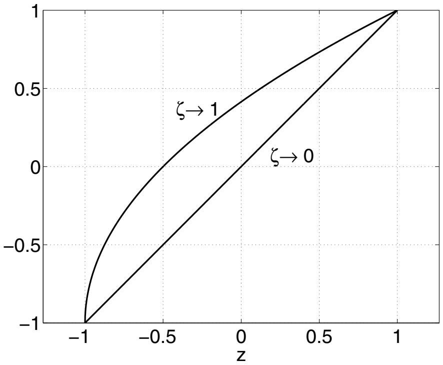


FIG. 5.1. *Variation of the outer temperature field for the limiting ζ values.*

The limit $\epsilon \to 0$ corresponds to the absence of a particle. Since the slip over the particle surface is the animating mechanism for the generation of flow, this limit represents a steady heat conduction process. Moreover, the transport problem becomes one-dimensional: $\theta = \tilde{\theta}(z)$. Thus, the energy equation adopts the form

$$\frac{d}{dz}\left[\left(1+\zeta\tilde{\theta}\right)\frac{d\tilde{\theta}}{dz}\right] = 0. \tag{5.1}$$

The solution to this equation, which satisfies the boundary conditions (3.10), is

$$\tilde{\theta}\left(z\right) = \frac{\sqrt{\zeta^2+2\zeta z+1}-1}{\zeta}. \tag{5.2}$$

For $\zeta \to 0$, $\tilde{\theta}\left(z\right) \sim z$; this is the linearized solution. For $\zeta \to 1^-$, $\tilde{\theta}\left(z\right) \sim \sqrt{2z+2}-1$. The dependence of $\tilde{\theta}$ upon ζ is depicted in Figure 5.1. Note that (5.2) implies that the temperature gradient is inversely proportional to the absolute temperature:

$$\frac{d\tilde{\theta}}{dz} = \frac{1}{1+\zeta\tilde{\theta}}. \tag{5.3}$$

Only in the limit $\zeta \to 0$ does the outer solution correspond to the notion of an "applied gradient."

Since the preceding solution does not satisfy the no-flux condition (3.11), it actually constitutes an outer solution of the limit $\epsilon \to 0$. Thus, the temperature distribution (5.2) (with vanishing $\boldsymbol{v}$ and $\tilde{p}$) needs to be supplemented by a comparable inner solution, which is valid for $r = O\left(\epsilon\right)$. When solving the thermophoretic problem using inner–outer expansions, separate asymptotic expansions in $\epsilon \ll 1$ are required in each separate region. These expansions must match in their common region of validity.

A solution comprised of a quiescent gas in the temperature profile (5.2) actually satisfies the exact governing equations. However, since it is not guaranteed to match the inner solution at all asymptotic orders, it constitutes only a leading-order outer solution. As will become evident in subsequent analysis, the leading-order inner solution consists of a temperature dipole and a velocity doublet. This solution "induces," through matching requirements, an $O\left(\epsilon^2\right)$ temperature correction and an $O\left(\epsilon^3\right)$ velocity correction in the outer solution. Luckily, the determination of such higher-order corrections proves unnecessary in the present investigation.

6. Inner problem. The inner problem is formulated in terms of the relative position vector. The rescaled variables of the inner problem are denoted by capital letters. Thus, the position and time variables are defined by $\boldsymbol{r} = \epsilon\boldsymbol{R}$ and $t = \epsilon T$, while the dependent variables are rescaled as

$$\Theta = \theta, \quad P = \epsilon\tilde{p}, \quad \boldsymbol{V} = \boldsymbol{v}, \quad \boldsymbol{W} = \boldsymbol{w}. \tag{6.1}$$

Note that the pressure rescaling is equivalent to normalization with the Stokes viscous scale $\mu_\infty\mathscr{U}/a$ (cf. (4.2)). The inner and outer coordinates are related via the relation

$$\boldsymbol{x} = \boldsymbol{x}_P(0) + \epsilon\int_0^T \boldsymbol{W}(\tau)\,d\tau + \epsilon\boldsymbol{R}, \tag{6.2}$$

and, specifically,

$$z = z_p(0) + \epsilon\int_0^T W(T')\,dT' + \epsilon R\cos\vartheta, \tag{6.3}$$

where $R = |\boldsymbol{R}|$ and with the polar angle ϑ measured from the z-axis. These two variables therefore constitute the radial and polar coordinates in a particle-fixed spherical reference system.

The inner temperature field is governed by the enthalpy equation (cf. (4.5))

$$\epsilon\frac{2\zeta}{3(1+\zeta\Theta)}\frac{D\Theta}{DT} = \nabla\cdot[(1+\zeta\Theta)\nabla\Theta], \tag{6.4}$$

together with the no-flux condition

$$\frac{\partial\Theta}{\partial R} = 0 \quad \text{at} \quad R = 1. \tag{6.5}$$

Here, $\nabla \equiv \partial/\partial\boldsymbol{R}$ is the inner gradient operator. The modified material derivative operator,

$$\frac{D}{DT} = \frac{\partial}{\partial T} + (\boldsymbol{V}-\boldsymbol{W})\cdot\nabla, \tag{6.6}$$

reflects the transformation (6.2) to a particle-fixed reference system. In addition to satisfying (6.4)–(6.5), Θ should also match the outer solution (5.2).

The inner flow field is governed by the differential equations (cf. (4.6)–(4.7)),

$$\nabla\cdot\boldsymbol{V} = \frac{\zeta}{1+\zeta\Theta}\frac{D\Theta}{DT}, \tag{6.7}$$

$$\epsilon\frac{\zeta}{1+\zeta\Theta}\frac{D\boldsymbol{V}}{DT} = -\nabla P + 2\nabla\cdot[(1+\zeta\Theta)\,\mathsf{E}], \tag{6.8}$$

where

$$\mathsf{E} = \frac{1}{2}\left[\nabla\boldsymbol{V} + (\nabla\boldsymbol{V})^\dagger\right] - \frac{1}{3}(\nabla\cdot\boldsymbol{V})\,\mathsf{I} \tag{6.9}$$

is the rescaled rate of strain tensor. It also satisfies the slip condition on the particle surface:

$$\boldsymbol{V} - \boldsymbol{W} = \epsilon^{-1}c_S(1+\zeta\Theta)\nabla\Theta \quad \text{at} \quad R = 1, \tag{6.10}$$

and is required to decay at large R. The flow problem is closed by the application of Newton's second law to the particle:

$$\oint d^2\boldsymbol{n}\,\boldsymbol{n}\cdot\left[-P\mathsf{I}+2\left(1+\zeta\Theta\right)\mathsf{E}\right]=\frac{4\pi}{3}\epsilon\zeta\sigma\frac{d\boldsymbol{W}}{dT}. \tag{6.11}$$

Here, $\boldsymbol{n}$ is an outward-pointing unit vector, normal to the particle surface, and $d^2\boldsymbol{n}$ is a differential solid angle about it.

7. Asymptotic matching.

A convenient procedure for performing inner–outer matching is provided by Van Dyke laws [22]:

(7.1) The m-term inner expansion of (the n-term outer expansion) =
The n-term outer expansion of (the m-term inner expansion).

Note that this is a statement of strict equality. Applying this law requires rewriting the left-hand side of (7.1) in terms of the outer variables (or, alternatively, rewriting the right-hand side of (7.1) in terms of the inner variables). The transformation between the inner and outer coordinates, however, involves the variable $\boldsymbol{W}$, which is itself expanded into an asymptotic series in ϵ. Thus, (7.1) cannot be satisfied unless one assumes a priori that the expansion of W terminates after a finite number of terms.

We therefore abandon Van Dyke's method in favor of the more general approach of intermediate variables [12]. We define the intermediate position vector $\boldsymbol{\xi}$:

$$\boldsymbol{r}=\epsilon^{\alpha}\boldsymbol{\xi}, \tag{7.2}$$

where $0<\alpha<1$. This vector is related to the inner coordinate through the relation $\boldsymbol{\xi}=\epsilon^{1-\alpha}\boldsymbol{R}$, and to the outer coordinate through the relation

$$\boldsymbol{X}=\boldsymbol{X}_p\left(0\right)+\epsilon\int_0^T\boldsymbol{W}\left(\tau\right)d\tau+\epsilon^{\alpha}\boldsymbol{\xi}. \tag{7.3}$$

The matching procedure requires that the inner and outer expansions possess a common domain of validity, in which ξ is $O(1)$. The more terms are required to be matched, the smaller this common domain becomes. It is therefore expected that matching of higher-order expansions would decrease the upper bound on α.

For future reference, we express the outer temperature field (5.2) in terms of the intermediate variable:

$$\begin{aligned}\tilde{\theta}\sim\tilde{\theta}_P&+\frac{\epsilon^{\alpha}}{1+\zeta\tilde{\theta}_P}\xi\cos\theta+\frac{\epsilon\int_0^T W\left(\tau\right)d\tau}{1+\zeta\tilde{\theta}_P}\\&-\frac{\epsilon^{2\alpha}\zeta}{2(1+\zeta\tilde{\theta}_P)^3}\xi^2\cos^2\vartheta-\frac{\epsilon^{1+\alpha}\zeta\int_0^T W\left(\tau\right)d\tau}{(1+\zeta\tilde{\theta}_P)^3}\xi\cos\vartheta+O\left(\epsilon^{3\alpha},\epsilon^2\right).\end{aligned} \tag{7.4}$$

Here,

$$\tilde{\theta}_P=\frac{\sqrt{\zeta^2+2\zeta z_P\left(0\right)+1}-1}{\zeta} \tag{7.5}$$

is the value of undisturbed temperature field (5.2) at the original position (at time $T=0$) of the particle center.

8. Asymptotic analysis in the inner region. We postulate the following asymptotic expansions: Inner solution

$$\Theta \sim \Theta_0 + \epsilon\Theta_1 + \epsilon^2\Theta + \cdots , \tag{8.1a}$$

$$\boldsymbol{V} \sim \boldsymbol{V}_0 + \epsilon\boldsymbol{V}_1 + \cdots , \tag{8.1b}$$

and

$$P \sim P_0 + \epsilon P_1 + \cdots , \tag{8.1c}$$

together with a comparable expansion E. In principle, these expansions induce the following expansion for $\boldsymbol{W} = \boldsymbol{e}_z W$:

$$W \sim W_0 + \epsilon W_1 + \cdots . \tag{8.1d}$$

8.1. Leading-order temperature field. The leading-order temperature field is governed by the equations

$$(1 + \zeta\Theta_0)\, \nabla^2\Theta_0 + \zeta\, (\nabla\Theta_0)^2 = 0, \tag{8.2a}$$

$$\frac{\partial\Theta_0}{\partial R} = 0 \quad \text{at} \quad R = 1, \tag{8.2b}$$

which possess a trivial constant solution. Matching with (7.4) readily yields

$$\Theta_0 \equiv \tilde{\theta}_P. \tag{8.3}$$

To calculate the leading-order velocity field the next term is also needed; see the slip condition (6.10). It is convenient to define

$$\bar{\Theta}_1 = (1 + \zeta\tilde{\theta}_P)\Theta_1. \tag{8.4a}$$

This scaled variable satisfies the following equations:

$$\nabla^2\bar{\Theta}_1 = 0, \tag{8.4b}$$

$$\frac{\partial\bar{\Theta}_1}{\partial R} = 0 \quad \text{at} \quad R = 1. \tag{8.4c}$$

Accordingly, it consists of spherical harmonics of the form

$$c^{(0)}\,(T) + \sum_{n=1}^{\infty} c^{(n)}\,(T) \left(R^n + \frac{n}{n+1} R^{-n-1} \right) P^{(n)}\,(\cos\vartheta)\,. \tag{8.5}$$

Here, $P^{(n)}$ is the Legendre Polynomial of degree n.[4] Note that the terms in (8.5) appear in pair-combinations which satisfy the no-flux condition (8.4c).

Matching with (7.4) reveals that Θ_1 does not possess any modes of $n > 1$. Accordingly,

$$\bar{\Theta}_1 = c_1^{(0)}\,(T) + c_1^{(1)}\,(T) \left(R + \frac{1}{2R^2} \right) \cos\vartheta. \tag{8.6}$$

[4] It is usually denoted by P_n; here we use an unconventional notation to avoid confusion with the various terms of the inner pressure expansion (8.1c).

In terms of the intermediate variable, the inner temperature field adopts the expansion

$$\Theta_0 + \epsilon\Theta_1 \sim \tilde{\theta}_P + \epsilon^{\alpha}\frac{c_1^{(1)}}{1+\zeta\tilde{\theta}_P}\xi\cos\theta + \epsilon\frac{c_1^{(0)}}{1+\zeta\tilde{\theta}_P} + O\left(\epsilon^{3-2\alpha}\right). \tag{8.7}$$

Matching with (7.4) then yields

$$c_1^{(1)} = 1, \quad c_1^{(0)} = \int_0^T W_0\left(\tau\right) d\tau. \tag{8.8}$$

The time-dependent term $c_1^{(0)}$ accounts for the slowly-varying temperature background observed in a particle-fixed reference system as the particle moves through the nonuniform temperature field.

8.2. Leading-order flow. The leading-order velocity field is governed by the following equation set:

$$\nabla\cdot\boldsymbol{V}_0 = 0, \tag{8.9a}$$

$$(1+\zeta\tilde{\theta}_P)\nabla^2\boldsymbol{V}_0 - \nabla P_0 = 0, \tag{8.9b}$$

$$\boldsymbol{V}_0 - \boldsymbol{W}_0 = c_S\nabla\bar{\Theta}_1 \quad \text{at} \quad R = 1, \tag{8.9c}$$

$$\boldsymbol{V}_1 \to \boldsymbol{0} \quad \text{as} \quad R\to\infty. \tag{8.9d}$$

In addition, (6.11) implies that the particle appears force-free at this asymptotic level.

Now, $\bar{\Theta}_1$ satisfies Laplace's equation, the no-flux condition, and the far-field behavior for large R:

$$\nabla\bar{\Theta}_1 \to \boldsymbol{e}_z. \tag{8.10}$$

Accordingly, it satisfies the same equations as would the electric potential in a thin-Debye-layer electrophoretic problem [1]. Moreover, the velocity field satisfies the same equations as those that apply to electrophoresis of a constant-zeta-potential particle. Thus, we can exploit Morrison's classical analysis [17] to obtain a solution to the present problem. The particle translates with the velocity

$$W_0 = -c_S \tag{8.11}$$

(corresponding to the Smoluchowski velocity in electrophoresis), the flow is irrotational,

$$\boldsymbol{V}_0 - \boldsymbol{W}_0 = c_S\nabla\bar{\Theta}_1, \tag{8.12}$$

and $P_0 = 0$. In view of (8.4b), it is readily verified that (8.12) trivially satisfies the flow equations (8.9a)–(8.9b). Moreover, using the Gauss theorem allows us to write the force delivered by $\boldsymbol{V}_0$ as an integral over any surface enclosing the particle, not necessarily its boundary. With the stress field decaying as R^{-4} (corresponding to the dipole term in (8.7)), it becomes evident that this force vanishes, as required.[5]

[5] These results actually hold for nonspherical particle shapes (in which case $\bar{\Theta}_1$ would not be given by (8.6), but would still decay as a dipole to leading order). When considering such particles, it is also necessary to demonstrate that they remain torque-free. This is again verified using the Gauss theorem and the fast decay of $\nabla\nabla\bar{\Theta}_1$.

The velocity (8.11) is the same as that obtained by Epstein [8]. This may appear surprising, since the outer temperature field deviates from the linear gradient of Epstein's small-ζ analysis. Indeed, the outer temperature field (5.2) appears at the particle-scale description (represented by (8.3) and (8.6)) as a uniform gradient superimposed upon a uniform temperature. This gradient is smaller by a factor $1+\zeta\tilde{\theta}_P$ from that appropriate to its $\zeta \to 0$ limit; see (5.3). Moreover, the absolute temperature is larger by that factor. Considering the slip condition (1.1), we then find that the direct effect of deviation from the small-ζ uniform gradient is to reduce the slip by the factor $(1+\zeta\tilde{\theta}_P)^2$. On the other hand, the kinematic viscosity increases by the same factor; see (4.3) and (4.8). Thus, Epstein's result is recovered due to a fortuitous cancellation of effects.

For future reference, we note that the leading-order rate-of-strain satisfies the relations

$$\mathsf{E}_0 = \nabla \boldsymbol{V}_0 = c_S \nabla\nabla\bar{\Theta}_1, \quad \nabla \cdot \mathsf{E}_0 = \mathbf{0}. \tag{8.13}$$

8.3. Convective-driven temperature correction. To analyze the perturbation to the leading-order flow, we need to evaluate Θ_2. It is convenient to define the variable

$$\bar{\Theta}_2 = \frac{(1+\zeta\tilde{\theta}_P)^3}{\zeta}\Theta_2, \tag{8.14a}$$

which is governed by the following equations:

$$\nabla^2\bar{\Theta}_2 = \left(\frac{2}{3}c_S - 1\right)(\nabla\Theta_1)^2 - \frac{2}{3}c_S, \tag{8.14b}$$

$$\frac{\partial\bar{\Theta}_2}{\partial R} = 0 \quad \text{at} \quad R = 1. \tag{8.14c}$$

A particular integral to (8.14b) is

$$\bar{\Theta}_{2,p} = -\frac{R^2}{6} + \left(\frac{2}{3}c_S - 1\right)\left[\frac{1}{24R^4} + \left(\frac{1}{3R} + \frac{1}{12R^4}\right)P^{(2)}(\cos\vartheta)\right]. \tag{8.15a}$$

This solution, however, does not satisfy the no-flux condition (8.14c). Thus, we add to (8.15a) the following terms, which, together with (8.15a), retain the condition

$$-\frac{1}{3R} - \left(\frac{2}{3}c_S - 1\right)\left[\frac{1}{6R} + \frac{2}{9R^3}P^{(2)}(\cos\vartheta)\right]. \tag{8.15b}$$

These terms satisfy the homogeneous counterpart of (8.14b).

The solution obtained is not unique: we can add to it an additional homogeneous solution, say $\bar{\Theta}_{2,h}$, consisting of spherical-harmonics pairs of the form (8.5) (recall that these pairs retain the no-flux condition). Since the inner solution must be matched with (7.4), only modes 0, 1, and 2 of the general solution (7.4) can be added. Thus, (8.15c)

$$\bar{\Theta}_{2,h} = c_2^{(0)}(T) + c_2^{(1)}(T)\left(R + \frac{1}{2R^2}\right)\cos\vartheta + c_2^{(2)}(T)\left(R^2 + \frac{2}{3R^3}\right)P^{(2)}(\cos\vartheta).$$

An asymptotic expansion for Θ in the intermediate domain is (cf. (8.7))

$$\tilde{\theta}_P + \epsilon^\alpha c_1^{(1)} \xi \cos\vartheta + \epsilon c_1^{(0)} + \frac{\epsilon^{2\alpha}\zeta}{(1+\zeta\tilde{\theta}_P)^3}\xi^2 \left[c_2^{(2)} P^{(2)}(\cos\vartheta) - \frac{1}{6}\right] + \frac{\epsilon^{1+\alpha}\zeta}{(1+\zeta\tilde{\theta}_P)^3} c_1^{(1)} \xi \cos\vartheta + O\left(\epsilon^{3-2\alpha}, \epsilon^2\right).$$

Matching with (7.4) in conjunction with (8.11) readily yields

$$c_2^{(2)} = -\frac{1}{3}, \quad c_2^{(1)} = c_S T. \tag{8.16}$$

(It is not necessary to obtain $c_2^{(0)}$.) For this matching process the intermediate domain must be modified to the range $0 < \alpha < 2/3$.

9. Leading-order flow correction. The leading-order flow correction $(\boldsymbol{V}_1, P_1)$ is governed by the mass equation

$$\nabla \cdot \boldsymbol{V}_1 = \frac{\zeta}{1+\zeta\tilde{\theta}_P}\frac{D\Theta_1}{DT} \tag{9.1a}$$

and momentum equation

$$\frac{\zeta}{1+\zeta\tilde{\theta}_P}\frac{D\boldsymbol{V}_0}{DT} = -\nabla P_1 + 2\nabla\cdot\left[(1+\zeta\tilde{\theta}_P)\mathsf{E}_1 + \zeta\Theta_1\mathsf{E}_0\right], \tag{9.1b}$$

together with the boundary conditions

$$\boldsymbol{V}_1 - \boldsymbol{W}_1 = c_S\left[(1+\zeta\tilde{\theta}_P)\nabla\Theta_2 + \zeta\Theta_1\nabla\Theta_1\right] \quad \text{at} \quad R = 1, \tag{9.1c}$$

$$\boldsymbol{V}_1 \to \boldsymbol{0} \quad \text{as} \quad R \to \infty. \tag{9.1d}$$

Since W_0 is constant, (6.11) implies that the particle is also force-free at the $O(\epsilon)$ asymptotic level:

$$\oint d^2\boldsymbol{n}\,\boldsymbol{n}\cdot\left\{-P_1\mathsf{I} + 2\left[(1+\zeta\tilde{\theta}_P)\mathsf{E}_1 + \zeta\Theta_1\mathsf{E}_0\right]\right\} = \boldsymbol{0}. \tag{9.1e}$$

9.1. A fictitious Stokes-type problem. Define

$$P_1 = (1+\zeta\tilde{\theta}_P)\bar{P}_1. \tag{9.2}$$

Making use of (8.12) and (8.13), the momentum-balance equation (9.1b) and the force-free condition (9.1e) are, respectively, rewritten as

$$\nabla^2\boldsymbol{V}_1 - \nabla\bar{P}_1 = -\zeta c_S(c_S - 2)\nabla\Theta_1\cdot\nabla\nabla\Theta_1 - \frac{1}{3}c_S\zeta\nabla\left[(\nabla\Theta_1)^2\right] \tag{9.3}$$

and

$$\oint d^2\boldsymbol{n}\,\boldsymbol{n}\cdot\left(-\bar{P}_1\mathsf{I} + 2\mathsf{E}_1\right) = -2c_S\zeta\oint d^2\boldsymbol{n}\,\boldsymbol{n}\cdot\Theta_1\nabla\nabla\Theta_1. \tag{9.4}$$

Note the Stokes operator appearing in (9.3) and the constant-viscosity structure of the stress-like object appearing in (9.4). Thus, we have transformed the flow problem for the variable-viscosity fluid into a comparable problem governing a fictitious constant-viscosity flow field $\big(\boldsymbol{V}_1, \bar{P}_1\big)$. The latter problem is driven by a mass-source distribution (the right-hand side of (9.1a)), velocity slip (the right-hand side of (9.1)), a body force distribution (the right-hand side of (9.3)), and an external force on the particle (the right-hand side of (9.4)).

The flow field $(\boldsymbol{V}_1, \bar{P}_1)$ is decomposed as follows:

$$\boldsymbol{V}_1 = \boldsymbol{V}_I + \boldsymbol{V}_{II} + \boldsymbol{V}_{III} + \boldsymbol{V}_{IV}, \tag{9.5a}$$

$$\bar{P}_1 = P_I + P_{II} + P_{III} + P_{IV}, \tag{9.5b}$$

wherein the subfields $\boldsymbol{V}_I$, $\boldsymbol{V}_{II}$, $\boldsymbol{V}_{III}$, and $\boldsymbol{V}_{IV}$ all decay at large R. The flow problem is, respectively, decomposed into four problems. The first problem satisfies

$$\nabla^2 \boldsymbol{V}_I = \text{right-hand side of (9.3)}, \tag{9.6a}$$

$$P_I = 0. \tag{9.6b}$$

The second problem satisfies

$$\nabla^2 \boldsymbol{V}_{II} = 0, \quad P_{II} = 0, \tag{9.7a}$$

$$\nabla \cdot \boldsymbol{V}_{II} = \text{right-hand side of (9.1a)} - \nabla \cdot \boldsymbol{V}_I. \tag{9.7b}$$

The third problem is governed by the Stokes equations, together with a slip-type boundary condition:

$$\nabla^2 \boldsymbol{V}_{III} - \nabla P_{III} = 0, \quad \nabla \cdot \boldsymbol{V}_{III} = 0, \tag{9.8a}$$

$$\boldsymbol{V}_{III} = \text{right-hand side of (9.1c)} - \boldsymbol{V}_I - \boldsymbol{V}_{II} \quad \text{at} \quad R = 1. \tag{9.8b}$$

The fourth problem is also governed by the Stokes equations, but with the boundary condition

$$\boldsymbol{V}_{IV} = \boldsymbol{W}_1 \quad \text{at} \quad R = 1. \tag{9.9}$$

It is identical to the problem governing the slow translation of a sphere at velocity $\boldsymbol{W}_1 = \boldsymbol{e}_z W_1$ relative to an otherwise quiescent fluid, wherein the no-slip boundary applies.

Our interest is not in the detailed fields, bur rather in their effect upon W_1. This velocity is determined from the condition that the total hydrodynamic force exerted upon the particle by the flow field $(\boldsymbol{V}_1, \bar{P}_1)$ is equal to the right-hand side of (9.4).

9.2. Symmetry arguments. It is convenient here to define a "flip-symmetry" property: We will say that a scalar field has this symmetry if it is invariant under the transformation $z \to -z$. Similarly, an axisymmetric vector field is flip-symmetric if its transverse component is flip-symmetric, while its axial component is antisymmetric. The latter definition can be equivalently stated using polar spherical coordinates: a vector field is flip-symmetric if its radial component is flip-symmetric, while its angular component is antisymmetric. It is readily verified that the Laplacian operator preserves flip-symmetry (or antisymmetry) when applied to either scalar or vector fields; the same is true for the divergence operator: the divergence of a flip-symmetric (antisymmetric) vector field is a flip-symmetric (antisymmetric) scalar field.

The right-hand side of (9.6a) is flip-symmetric; accordingly, so is $\boldsymbol{V}_I$. Clearly, then, $\boldsymbol{V}_I$ does not contribute to the hydrodynamic on the particle. Moreover, it is readily seen that the right-hand side of (9.7b) is flip-symmetric, whence so is $\boldsymbol{V}_{II}$, which then also does not contribute to the hydrodynamic force.

Consider now the slip-driven field $\boldsymbol{V}_{III}$. The slip-terms in the right-hand-side of (9.8b) which are contributed by $\boldsymbol{V}_I$ and $\boldsymbol{V}_{II}$ are flip-symmetric. They therefore contribute flip-symmetric components to $\boldsymbol{V}_{III}$; these components, obviously, do not generate any hydrodynamic force. Moreover, the antisymmetric terms in the right-hand side of (9.1c) cancel each other. We conclude that $\boldsymbol{V}_{III}$, too, does not contribute to the force.

The hydrodynamic force contributed by the fourth hydrodynamic subfield is simply the Stokes drag $-6\pi\boldsymbol{W}_1$. Thus, W_1 is determined by equating this force to the right-hand side of (9.4). Straightforward calculation shows that the latter vanishes. Accordingly, $W_1 = 0$.

10. Maxwell stresses. From the viewpoint of gas-kinetic theory, viscous stresses (and the heat-flux vector) constitute $O\left(\mathrm{Kn}\right)$ correction terms in the Chapman–Enskog expansion of the Boltzmann equation [6], wherein the ideal-fluid model emerges at the leading-order. Higher-order terms obtained in that expansion, such as the $O\left(\mathrm{Kn}^2\right)$ Burnett terms, are traditionally not included in conventional continuum descriptions.

The exclusion of Burnett stresses, however, has been shown to be inconsistent in flows driven by significant temperature differences [9, 13]. The inconsistency is related to only two of the six Burnett stress terms, which are provoked by temperature gradients. For Maxwell molecules, these two "thermal" terms adopt the dimensional form

$$-3\frac{\mu^{*2}}{\rho^*\theta^*}\frac{\partial^2\theta^*}{\partial\boldsymbol{x}^*\partial\boldsymbol{x}^*}, \quad -3\frac{\mu^{*2}}{\rho^*\theta^{*2}}\frac{\partial\theta^*}{\partial\boldsymbol{x}^*}\frac{\partial\theta^*}{\partial\boldsymbol{x}^*}. \tag{10.1}$$

In the presence of a single length scale L, these terms are of order $\zeta{\mu^*_\infty}^2/\rho^*_\infty L^2$. The ratio of these stresses to the $O\left(\mu^*_\infty\mathscr{U}/L\right)$ viscous terms (where $\mathscr{U}$ is a characteristic velocity of the flow) therefore scale as $\zeta\nu^*_\infty/L\mathscr{U}$. Since the kinematic viscosity ν^*_∞ is $O\left(\lambda c^*_\infty\right)$, where λ is the mean free path, this ratio is about $\zeta\,\mathrm{Kn}/M$, and is indeed $O\left(\mathrm{Kn}\right)$ small for "normal" ($M = O\left(1\right)$) situations considered in gas-kinetic theory.[6]

For slip-driven flows, however, the velocity $\mathscr{U}$ scales according to (1.3), and the above-mentioned ratio is actually $O\left(1\right)$: the thermal stresses are of comparable magnitude to the viscous (and inertial) terms. It was shown in [9] that the remaining Burnett terms are still $O\left(\mathrm{Kn}\right)$ small, and are therefore negligible.

In the present dimensionless momentum equation, the right-hand side of (4.7) is accordingly supplemented by the term

$$-3\frac{\partial}{\partial\boldsymbol{x}}\cdot\left[\left(1+\zeta\theta\right)^2\frac{\partial^2\theta}{\partial\boldsymbol{x}\,\partial\boldsymbol{x}} + \zeta\left(1+\zeta\theta\right)^2\frac{\partial\theta}{\partial\boldsymbol{x}}\frac{\partial\theta}{\partial\boldsymbol{x}}\right], \tag{10.2}$$

which corresponds to the divergence of the Maxwell stresses (10.1).

In principle, the existence of Maxwell stresses implies that classical pure-conduction situations may not be compatible with mechanical equilibrium. In the present investigations, it is therefore necessary to reexamine the validity of the outer solution. Fortunately, the one-dimensional temperature profile (5.2) generates Maxwell stresses which possess only an $\boldsymbol{e}_z\boldsymbol{e}_z$ component:

$$-3\left[\left(1+\zeta\theta\right)^2\frac{d^2\theta}{dz^2} + \zeta\left(1+\zeta\theta\right)^2\left(\frac{d\theta}{dz}\right)^2\right].$$

[6] Usually it is even smaller, since in the absence of *imposed* temperature differences, ζ is also small; if the temperature field is set by viscous dissipation, for example, it is $O\left(\mathrm{Kn}\right)$.

According to (5.3), this component vanishes.

Of course, the Maxwell stresses do modify the inner solution. It is readily verified that the right-hand side of the inner momentum equation (6.8) is supplemented by the term

$$-3\epsilon^{-1}\nabla\cdot\left[(1+\zeta\Theta)^2\,\nabla\nabla\Theta+\zeta\,(1+\zeta\Theta)\,\nabla\Theta\nabla\Theta\right]. \tag{10.3}$$

In the leading-order flow problem, the right-hand side of (8.9b) is supplemented by the term

$$-3\nabla\cdot\left[(1+\zeta\tilde{\theta}_P)^2\nabla\nabla\Theta_1\right].$$

Since Θ_1 is harmonic (see (8.4b)), this term vanishes. Consider now the leading-order flow correction: the right-hand side of (9.3) is supplemented by the term

$$-3\nabla\cdot\left[(1+\zeta\tilde{\theta}_P)\nabla\nabla\Theta_2+2\zeta\Theta_1\nabla\nabla\Theta_1+\zeta(1+\zeta\tilde{\theta}_P)\nabla\Theta_1\nabla\Theta_1\right].$$

It is easily verified that this term is flip-symmetric and does not contribute to the hydrodynamic force experienced by the particle.

11. Concluding remarks.

The drift of a spherical particle in a nonisothermal gaseous environment, provoked by a thermally-induced slip at the outer edge of a Knudsen layer, was investigated theoretically. The gas is bounded between two parallel walls maintained at uniform but unequal temperatures. Unlike previous studies, it was not assumed that the temperature difference is small compared with the mean absolute temperature. The problem is then inherently nonlinear, with a universal scaling for the Reynolds and Péclet numbers.

With the goal of obtaining qualitative understanding, we adopt the simple model of an insulating particle. (At the other extreme, when the particle conductivity is large compared with that of the gas, the slip model obviously predicts null thermophoretic velocity.) We also restrict the analysis to the continuum regime, the Knudsen number approaching zero.

The analysis is asymptotic, in the limit of a small particle. In view of the inherent nonlinearity, it is impossible to transform the problem into a steady one. The small-particle limit is singular and requires asymptotic matching between the particle-scale solution and the apparatus-scale nonlinear behavior. The transformation between the inner and outer descriptions employs the particle position, itself a dependent variable of the asymptotic problem.

Due to a fortuitous cancellation of terms, we found that Epstein's result [8] holds at leading order. Moreover, the symmetric leading-order $O(\epsilon)$ correction to the flow field does not affect Epstein's result. The inclusion of Burnett stresses modifies the flow field, but symmetry arguments show that this modification does not contribute to the thermophoretic velocity at the inspected asymptotic orders.

In view of the growing interest in the effect of walls upon thermophoretic motion [7, 11] and the $O(\epsilon^3)$ asymptotically small wall effects predicted by prevailing analyses of the linearized model [1], it is desirable to evaluate higher-order corrections to the flow. Unfortunately, we were unable to progress further with the present asymptotic scheme. Thorough understanding of the nonlinear thermophoretic mechanism may therefore require analysis of the nonlinear model in the future.

REFERENCES

[1] J. L. Anderson, *Colloid transport by interfacial forces*, in Annu. Rev. Fluid Mech., 30 (1989), pp. 139–165.

[2] B. K. Annis, *Thermal creep in gases*, J. Chem. Phys., 57 (1972), pp. 2898–2905.

[3] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 1967.

[4] A. V. Bobylev, *Quasistationary hydrodynamics for the Boltzmann equation*, J. Stat. Phys., 80 (1995), pp. 1063–1083.

[5] I. D. Chang, *Slow motion of a sphere in a compressible viscous fluid*, Z. Angew. Math. Phys., 16 (1965), pp. 449–469.

[6] S. Chapman and T. G. Cowling, *The Mathematical Theory of Non-Uniform Gases*, 3rd ed., Cambridge University Press, London, 1970.

[7] S. H. Chen, *Boundary effects on a thermophoretic sphere in an arbitrary direction of a plane surface*, AIChE J., 46 (2000), pp. 2352–2368.

[8] P. S. Epstein, *Zur Theorie des Radiometers*, Z. Physik., 54 (1929), pp. 537–563.

[9] V. S. Galkin, M. N. Kogan, and O. G. Fridlender, *Obtekanie sil'no nagretoi sfery potokom gaza pri malykh chislakh Reinol'dsa*, Prikl. Mat. Mekh., 5 (1972), pp. 880–885.

[10] D. R. Kassoy, T. C. Adamson, and A. F. Messiter, *Compressible low Reynolds number flow around a sphere*, Phys. Fluids, 9 (1966), pp. 671–681.

[11] H. J. Keh and P. Y. Chen, *Thermophoresis of an aerosol sphere parallel to one or two plane walls*, AIChE J., 49 (2003), pp. 2283–2299.

[12] J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods*, Springer, New York, 1996.

[13] M. N. Kogan, *Molecular gas dynamics*, Ann. Rev. Fluid Mech., 5 (1973), pp. 383–403.

[14] M. N. Kogan, V. S. Galkin, and O. G. Fridlender, *Stresses produced in gasses by temperature and concentration inhomogeneities. New types of free convection*, Sov. Phys. Usp., 19 (1976), pp. 420–428.

[15] J. M. Los and R. R. Fergusson, *Measurements of thermomolecular pressure differences on argon and nitrogen*, Trans. Faraday Soc., (1952), pp. 730–738.

[16] J. C. Maxwell, *On stresses in rarefied gases resulting from inequalities of temperature*, Philos. Trans. R. Soc. Lond. Ser. A, 170 (1879), pp. 231–262.

[17] F. A. Morrison, *Electrophoresis of a particle of arbitrary shape*, J. Colloid Interface Sci., 34 (1970), pp. 210–214.

[18] D. H. Papadopoulos and D. E. Rosner, *Enclosure gas-flows driven by nonisothermal walls*, Phys. Fluids, 7 (1995), pp. 2535–2537.

[19] O. Reynolds, *On certain dimensional properties of matter in the gaseous state. Part* I. *Experimental researches on thermal transpiration of gases through porous plates and on the laws of transpiration and impulsion, including an experimental proof that gas is not a continuous plenum. Part* II. *On an extension of the dynamical theory of gas, which includes the stresses, tangential and normal, caused by a varying condition of gas, and affords an explanation of the phenomena of transpiration and impulsion*, Philos. Trans. R. Soc. Lond. Ser. A, 170 (1879), pp. 725–845.

[20] Y. Sone, *Thermal creep in rarefied gas*, J. Phys. Soc. Japan, 21 (1966), pp. 1836–1837.

[21] J. Tyndall, *On dust and disease*, Proc. R. Inst., 6 (1870), pp. 1–14.

[22] M. Van Dyke, *Perturbation Methods in Fluid Mechanics*, Academic Press, New York, 1964.

[23] E. Yariv, *Slip-driven thermal rectification*, Europhys. Lett. EPL, 79 (2007), p. 24001.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 473–493

OSTWALD RIPENING IN THIN FILM EQUATIONS[*]

K. B. GLASNER[†]

Abstract. Fourth order thin film equations can have late stage dynamics that are analogous to the classical Cahn–Hilliard equation. We undertake a systematic asymptotic analysis of a class of equations that describe partial wetting with a stable precursor film introduced by intermolecular interactions. The limit of small precursor film thickness is considered, leading to explicit expressions for the late stage dynamics of droplets. Our main finding is that exchange of mass between droplets characteristic of traditional Ostwald ripening is a subdominant effect over a wide range of kinetic exponents. Instead, droplets migrate in response to variations of the precursor film. Timescales for these processes are computed using an effective medium approximation to the reduced free boundary problem, and dynamic scaling in the reduced system is demonstrated.

Key words. thin film equation, dewetting, coarsening, Ostwald ripening

AMS subject classifications. 76D08, 34E05

DOI. 10.1137/080713732

1. Introduction.

Viscous liquid films have a rich set of dynamics that are still only partially understood [7, 28]. A large subset of phenomena involves dewetting instabilities that produce a diverse collection of patterns that have been studied experimentally [1, 13, 33, 34, 35, 42] as well as theoretically [2, 3, 19, 22, 27, 36, 37, 42]. These instabilities cause nearly uniform fluid layers to break up into arrays of large droplets connected by a remaining (very) thin film, which undergo an elaborate coarsening process characterized both by coalescence of droplets and exchange of fluid between droplets and the intervening film [11, 12, 15, 16].

The results we describe run parallel to other studies of dynamical coarsening processes, most notably phase separation phenomena described by the Cahn–Hilliard equation [5, 30]. At late times and small volume fractions, this equation describes the Ostwald ripening process [14, 24, 25, 38, 39]. Our purpose is to describe a similar limit for a class of thin film equations and to highlight the differences between our problem and classical Ostwald ripening.

The analogy between spinodal decomposition and liquid dewetting was first explored by Mitlin [19, 20, 21, 22]. Subsequent theoretical works have studied coarsening in thin film equations that results from other instabilities. Bestehorn, Pototsky, and Thiele [4] consider the evolution of a film destabilized by Marangoni effects and quantify coarsening rates through numerical experiments. Merkt et al. [18] obtain similar results for a two-layer film. There are, in fact, numerous other variations of dissipative fourth order equations that exhibit coarsening behavior, for example, the convective Cahn–Hilliard equation studied by Watson et al. [40].

This paper is a continuation of a body of work initiated by Glasner and Witelski [11, 12] on coarsening behavior of liquid droplets described by disjoining-pressure models. It was found there that the dewetting instability leads to the eventual development of droplets separated by a precursor film. The subsequent one-dimensional

[*] Received by the editors January 18, 2008; accepted for publication (in revised form) July 18, 2008; published electronically November 14, 2008. This work was produced with the assistance of NSF award DMS-0405596.

http://www.siam.org/journals/siap/69-2/71373.html

[†] Department of Mathematics, University of Arizona, 617 N. Santa Rita, Tucson, AZ 85721 (kglasner@math.arizona.edu).

dynamics of these droplets was computed, involving mass exchange between droplets and the precursor layer as well as motion of the droplets themselves. This results in a coarsening process characterized by both mass transfer and coalescence, and exhibits dynamic scaling with a nonstandard exponent. Rigorous bounds for dynamic scaling were subsequently obtained by Otto, Rump, and Slepčev [29]. In two dimensions, the interaction of droplets has been studied by Pismen and Pomeau [32]. Although not entirely dissimilar from the conclusions described here, their results are in both quantitative and qualitative disagreement with our calculations (see the concluding section for a comparison).

This work serves as a companion paper to the manuscript of Glasner et al. [9]. Instead of a matched asymptotics approach, that work utilizes a variational principle (the Rayleigh–Onsager notion of least dissipation [26]) to explain and quantify droplet migration effects. Both papers obtain comparable results, although a careful interpretation is needed to show their equivalence. Some comparison is provided in section 5.

This paper considers a class of fourth order parabolic equations which have the structure

$$\tau(\epsilon)h_t = \nabla\cdot(h^q\nabla p), \qquad p = \epsilon^{-1}U'\left(\frac{h}{\epsilon}\right) - \Delta h, \quad q > 0. \tag{1.1}$$

The physical domain is taken to be a two-dimensional, bounded, simply connected open set Ω, where Neumann and no-flux boundary conditions are imposed (although few of our results depend crucially on these assumptions). The timescale $\tau(\epsilon)$ is chosen to capture the slow dynamics associated with migration and mass exchange (i.e., ripening) of droplets. It depends on the mobility exponent as

$$\tau(\epsilon) = \begin{cases} \epsilon^q, & q \in (0,2), \\ \epsilon^2 \ln(\epsilon^{-1}), & q = 2, \\ \epsilon^2, & q \in (2,3), \\ \epsilon^2/\ln(\epsilon^{-1}), & q = 3, \\ \epsilon^{q-1}, & q > 3. \end{cases} \tag{1.2}$$

Our interest is in the limit of small ϵ, which corresponds to both thin precursor films and long timescales.

It is instructive to consider a range of mobility exponents to capture the crossover between different dynamical mechanisms. For liquid films, this exponent is a function of the solid-liquid boundary condition and the fluid rheology. The standard case of a Newtonian fluid with a no-slip boundary condition corresponds to $q = 3$. The Navier slip condition leads to $q = 2$ if particular limits are considered [8, 23], whereas Darcy's law can lead to $q = 1$ [6]. Models of non-Newtonian fluids may have a variety of exponents (see, e.g., [41]).

The class of potentials U considered here include those commonly employed to describe a combination of attractive and repulsive van der Waals forces [28]. The following assumptions are used:

1. U is scaled so that it has a minimum at 1 and $U(\infty) - U(1) = 2$.
2. U' has a unique maximum at $H^* > 1$.
3. The potential decays as

$$U'(H) = \mathcal{O}(H^{-\alpha}), \quad H \to \infty, \tag{1.3}$$

where

$$\alpha > \begin{cases} q+1, & q \in (0,2), \\ 3, & q \geq 2. \end{cases} \tag{1.4}$$

This will ensure that intermolecular interactions have a subdominant effect for macroscopic ($h \sim \mathcal{O}(1)$) films.

The structure of the paper is as follows. Section 2 describes the results of the lengthy calculation, whose details are given in sections 3 and 4. Section 5 goes on to propose an approximation procedure for the resulting free boundary problem, and timescales for the relevant dynamics are worked out. Section 6 gives example calculations and compares them to predictions of dynamic scaling.

2. Setup for matched asymptotics and a summary of results. There will be three regions in the matched asymptotic analysis (see Figure 2.1):

- Region I: This region corresponds to droplets and is composed of the union of disjoint disks $\{D_i\}$ which have the form $D = \{\mathbf{x} \,\big|\, |\mathbf{x} - X| < R\}$ so that X is the droplet center and R is its radius. Unit normals to ∂D will be denoted $\mathbf{n}$, and we will also utilize the coordinate unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, etc. In this region, h and x will both scale like $\mathcal{O}(1)$. It will be convenient to use the moving polar coordinates $r = |\mathbf{x} - X(t)|$, $\theta = \arg(\mathbf{x} - X(t))$.
  To be more precise about R and X, we define the contact line at finite ϵ to be the set $\{\mathbf{x} \mid h(\mathbf{x}) = \epsilon H^*\}$, where H^* is the global maximum of U'. This definition is somewhat arbitrary and is chosen merely for convenience. On the other hand, in the limit $\epsilon \to 0$ this set converges to the boundary of the support of h, i.e., the sharp-interface contact line. We suppose that for each droplet this curve is nearly circular and has the form $\mathbf{x} = X + R(\theta)\hat{\mathbf{r}}$. Properly speaking, R and X should also be expanded in ϵ, but to avoid

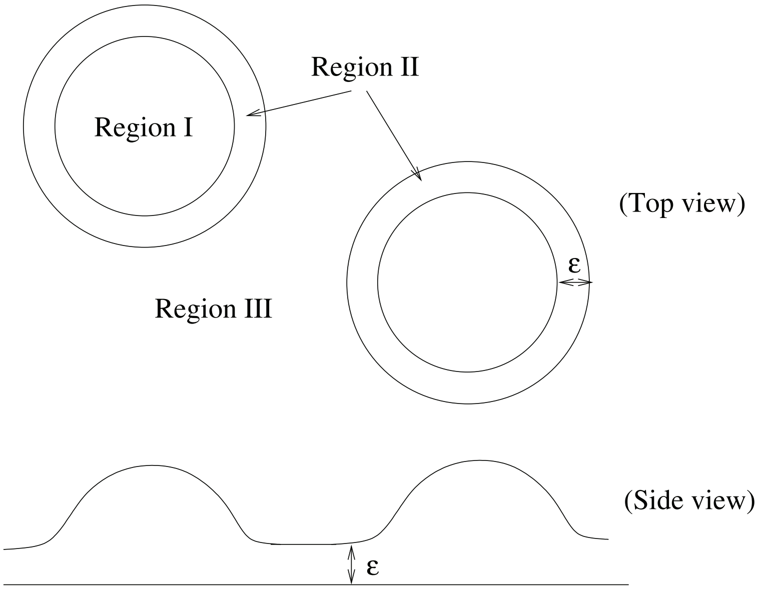


FIG. 2.1. *Three regions for the matched asymptotic calculation.*

excessive notation, these labels will simply denote the corresponding leading order solutions. In particular, we find that R is independent of θ at leading order.

- Region II is a microscopic internal layer near the contact line where h and x scale like ϵ. The moving rescaled radial coordinate

$$z = \frac{R(t) - r}{\epsilon} \tag{2.1}$$

will be employed. In light of the definition of R, the solution in this region must satisfy

$$h(z = 0) = h^*. \tag{2.2}$$

For reference, the Laplacian in z, θ coordinates expands as

$$\Delta h = \epsilon^{-2} h_{zz} - \epsilon^{-1} R^{-1} h_z - \Big(z R^{-2} h_z + R^{-2} h_{\theta\theta} \Big) + \mathcal{O}(\epsilon). \tag{2.3}$$

- Region III is the complement $\Omega / \cup D_i$ which contains the precursor film. In this region, h will scale like ϵ.

The overall strategy is to propose self-consistent asymptotic expansions in each region and to connect them via matching conditions. Less-standard matching conditions are derived when needed. Corrections to the leading order base solutions solve linear equations, and Fredholm-type solvability conditions will yield information about the dynamics.

The main goal is to determine the dynamic behavior of R and X, which will be shown to arise from a flux J which is determined by the elliptic problem

$$\Delta P = 0, \quad P|_{\partial D_i} = \frac{2}{R_i}, \quad J = -\nabla P, \tag{2.4}$$

solved in the exterior region $\Omega / \cup D_i$. Here P represents the first nontrivial correction to the pressure p. Equations (2.4) describe quasi-steady diffusion of material driven by a Gibbs–Thomson boundary condition. We find that, with respect to the timescale τ, the dynamics at leading order are

$$R_t = \begin{cases} -\frac{4}{3\pi R^2} \int_{\partial D} J \cdot \mathbf{n} \, ds, & q < 2, \\ 0, & q \geq 2, \end{cases} \tag{2.5}$$

and

$$X_t = -M(R) \int_{\partial D} \mathbf{n} J \cdot \mathbf{n} \, ds, \tag{2.6}$$

where the mobility factor $M(R)$ is

$$M = \frac{1}{\pi} \begin{cases} R^{-2} \Psi_1(1) \Big/ \int_0^1 \Psi_1(r) r^2 dr, & q < 2, \\ R^{-2} \Big/ \int_0^1 \Psi_1(r) r^2 dr, & q = 2, \\ R^{q-4} \int_{-\infty}^{\infty} [H_1^{1-q} - H_1^{-q}] dz \Big/ \int_0^1 \Psi_1(r) r^2 dr, & q \in (2,3), \\ R^{-1} \int_{-\infty}^{\infty} [H_1^{-2} - H_1^{-3}] dz, & q = 3, \\ R^{-1} \int_{-\infty}^{\infty} [H_1^{1-q} - H_1^{-q}] dz \Big/ \int_{-\infty}^{\infty} H_1^{2-q} - H_1^{1-q} dz, & q > 3. \end{cases} \tag{2.7}$$

H_1 is the leading order solution for the microscopic contact line region II. The function Ψ_1 arises from the solvability argument and is specified as the solution of the rescaled boundary value problem (4.5)–(4.7).

The point of writing $R_t = 0$ for $q \geq 2$ in (2.5) is to emphasize the crossover between radial and migration dynamics. This is to some extent an artifact of our choice of timescales (1.2). Had we chosen $\tau = \epsilon^q$ for all q instead, then the radial dynamics would just be given by the first nonzero formula in (2.5).

3. Base solutions.

This section summarizes the aspects of the analysis which are common to all mobility exponents $q > 0$. The rest is split into cases in the following section.

Region II. The solution is expanded as $h = \epsilon H_1 + \epsilon^2 H_2 + o(\epsilon^2)$. The leading order equation is

$$(H_1^q[-(H_1)_{zz} + U'(H_1)]_z)_z = 0. \tag{3.1}$$

Integrating twice and using the matching conditions $(H_1)_z \sim 0$ as $z \to -\infty$, we get

$$-(H_1)_{zz} + U'(H_0) = c_1. \tag{3.2}$$

The matching condition $(H_1)_{zz} \sim 0$ as $z \to +\infty$ means that $C = 0$ in light of (1.3). It follows that $H_1 \sim 1$ as $z \to -\infty$, and we can integrate again to obtain

$$\frac{1}{2}(H_1)_z^2 = U(H_1) - U(1), \tag{3.3}$$

from which the equilibrium contact angle is determined by

$$(H_1)_z = \sqrt{2[U(H_1) - U(1)]} \sim 1, \quad z \to +\infty. \tag{3.4}$$

Solving (3.3) gives the solution implicitly as

$$\int^{H_1} \frac{dH}{\sqrt{2[U(H) - U(1)]}} = z + c_2. \tag{3.5}$$

The constant of integration is determined uniquely by the condition (2.2).

The next order correction satisfies

$$(H_1^q[-(H_2)_{zz} - R^{-1}(H_1)_z + U''(H_1)H_2]_z)_z = 0. \tag{3.6}$$

Integrating and using the matching condition $(H_2)_{zzz} \to 0$ as $z \to \infty$ gives

$$[-(H_2)_{zz} - R^{-1}(H_1)_z + U''(H_1)H_2]_z = 0. \tag{3.7}$$

A further integration implies

$$-(H_2)_{zz} - R^{-1}(H_1)_z + U''(H_1)H_2 \equiv P = \text{constant}. \tag{3.8}$$

This says that (total) leading order pressure is constant through region II, and we can use this to match between regions I and III. We remark that both H_1 and H_2 are independent of θ. Later in the calculation, this will provide symmetry that is needed to make certain integrals vanish.

Region I. Expanding $h = h_0(\mathbf{x}, t) + o(1)$ for now, we obtain

$$\nabla\cdot(h_0^q \nabla \Delta h_0) = 0, \quad \mathbf{x} \in D. \tag{3.9}$$

Provided that h_0 is well behaved (bounded third derivatives), integration of (3.9) against Δh_0 gives

$$\int_D h_0^q |\nabla \Delta h_0|^2 dx = 0. \tag{3.10}$$

Since $h_0 \to 0$ on the boundary of D, it follows that Δh_0 is a constant. Using the matching conditions

$$h_0(R,\theta) = 0, \quad (h_0)_r(R,\theta) = -1 \tag{3.11}$$

gives the family of radially symmetric droplet solutions

$$h_0(\mathbf{x}; R(t), X(t)) = R(t) H\left(\frac{\mathbf{x} - X(t)}{R(t)}\right), \quad H(\eta) = \frac{1}{2}(1-\eta^2). \tag{3.12}$$

Using (3.8), (3.12) and the matching condition

$$(h_0)_{rr}(R,\theta) = \lim_{z\to\infty} (H_2)_{zz} \tag{3.13}$$

allows us to relate the pressure P and the droplet radius:

$$P = -\Delta h_0 = \frac{2}{R(t)}. \tag{3.14}$$

Region III. Here we expand $h = \epsilon h_1 + \epsilon^2 h_2 + o(\epsilon^2)$. Because of the scaling of $\tau(\epsilon)$, the leading order problem for all $q > 0$ is the elliptic equation

$$\nabla \cdot (h_1^q \nabla U'(h_1)) = 0. \tag{3.15}$$

Matching to region II implies $h_1 = 1$ on the boundary $\cup \partial D_i$; therefore $h_1 \equiv 1$. At order ϵ^2, the correction term satisfies the "quasi-steady" problem

$$\Delta h_2 = 0. \tag{3.16}$$

This equation is solved together with boundary conditions that are derived by matching. Using (3.8) and (3.14), we find that

$$U''(1) h_2 = \frac{2}{R(t)}, \quad \mathbf{x} \in \partial D. \tag{3.17}$$

It is convenient introduce the flux

$$J = -h^q \nabla p = -\epsilon^q U''(1) \nabla h_2 + o(\epsilon^q) \tag{3.18}$$

so that at leading order

$$J_q = -\nabla P, \quad P \equiv U''(1) \nabla h_2 \tag{3.19}$$

is therefore determined by solving the boundary value problem (3.16), (3.17). To avoid excessive notation, we also use J_q to denote the flux of order $\mathcal{O}(\epsilon^q)$ in regions I and II.

4. Mobility-dependent expansions.

4.1. Case $q \in (0, 2)$. The expansion of the equation in region II at order ϵ^q gives $0 = (J_q \cdot \hat{\mathbf{z}})_z$, which merely says that the z-component of J_q is constant through this layer. Thus the normal component of J_q to the boundary of D is that given by the solution in region III.

In region I, we expand $h = h_0(x,t) + \epsilon^q h_q + o(\epsilon^q)$, which means that leading order flux is $J_q = h_0^q \nabla \Delta h_q$. The first nontrivial correction to the equation in this region gives the linear problem

$$\mathcal{L} h_q = X_t \cdot \nabla h_0 - R_t \frac{\partial h_0}{\partial R}, \quad \mathcal{L} = \nabla \cdot [h_0^q \nabla \Delta \tag{4.1}$$

for $\mathbf{x} \in D$ (the mismatched bracket indicates that the divergence applies to everything to the right). We remark that a similar linear problem was encountered by Pismen [31]. The linear operator $\mathcal{L}$ (on a space endowed with suitable homogeneous boundary conditions) has the adjoint

$$\mathcal{L}^\dagger = \Delta \nabla \cdot [h_0^q \nabla. \tag{4.2}$$

Nullspace of $\mathcal{L}^\dagger$. To invoke a Fredholm solvability argument, we need to characterize its nullspace by finding orthogonal functions whose span is the same as $\{(h_0)_x, (h_0)_y, (h_0)_R\}$. Observe that if ψ is in the nullspace, then

$$\nabla \cdot \left[h_0^q \nabla \psi \right] = \phi, \quad \Delta \phi = 0. \tag{4.3}$$

We shall be interested in the particular harmonic functions $\phi = 0, -x, -y$, which ultimately correspond to changes in droplet size and translation in each direction, respectively. Since $x = r\cos\theta$, $y = r\sin\theta$, we look for a solution of (4.3a) of the form $\psi = \Psi(r)\cos\theta$ or $\psi = \Psi(r)\sin\theta$. In either case we are led to the differential equation

$$r(r h_0^q \Psi')' - h_0^q \Psi = -r^3 \tag{4.4}$$

together with the boundary conditions

$$h_0^q \Psi'(R) = 0, \quad \Psi(0) = 0. \tag{4.5}$$

Several observations about (4.4)–(4.5) are in order. First, the solution is unique, since multiplying the homogeneous version of this linear equation by Ψ/r and integrating leads to

$$\int_0^R \left[r h_0^q \Psi'^2 + \frac{h_0^q \Psi^2}{r} \right] dr = 0. \tag{4.6}$$

There is also a natural scale invariance for this problem: If Ψ_1 solves

$$r(r H^q \Psi_1')' - H^q \Psi_1 = r^3, \quad \Psi_1(0) = 0, \quad H^q \Psi_1(1) = 0, \tag{4.7}$$

then

$$\Psi = R^{3-q} \Psi_1(r/R) \tag{4.8}$$

solves (4.4). Finally, the regularity of solutions of the ordinary differential equation (4.4) and the first boundary condition (4.5) allow us to ascertain the asymptotics at

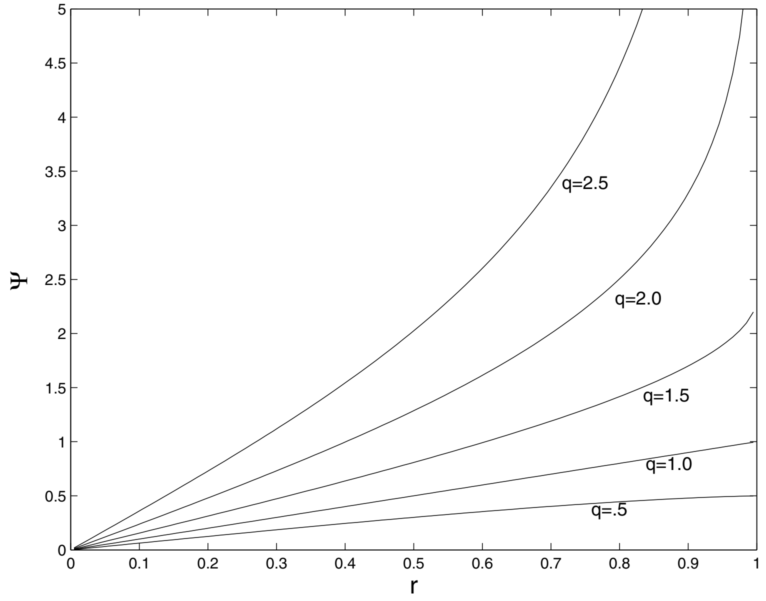


FIG. 4.1. *The function $\Psi(r)$ with $R = 1$, used in the solvability argument.*

$r = R$. In particular, we have $h_0^q\Psi'(R) = \mathcal{O}(|r - R|)$, and therefore one computes for $r \to R$

$$\Psi \sim \begin{cases} \mathcal{O}(1), & q < 2, \\ R\ln|r - R|, & q = 2, \\ \frac{R}{q-2}|r - R|^{2-q}, & q > 2. \end{cases} \tag{4.9}$$

In particular, Ψ is bounded for $q < 2$ and integrable for $q < 3$. Since $r = R$ is a regular singular point of (4.4), the first boundary condition in (4.5) implies

$$h_0^q\Psi'(r) = \mathcal{O}(|r - R|), \quad r \to R. \tag{4.10}$$

In practice, solutions to (4.4) can be obtained numerically (see Figure 4.1). To summarize, the desired functions for the solvability argument are

$$\psi_R = 1, \quad \psi_x = \Psi(r)\cos\theta, \quad \psi_y = \Psi(r)\sin\theta. \tag{4.11}$$

Solvability conditions. The inner product of ψ_R with (4.1) produces

$$R_t = -\frac{\int_{\partial D} h_0^q\nabla\Delta h_q \cdot \mathbf{n}\,ds}{\int_D \partial h_0/\partial R\,dx}. \tag{4.12}$$

Using the matching condition for flux,

$$R_t = -\frac{4}{3\pi R^2}\int_{\partial D} J_q \cdot \mathbf{n}\,ds. \tag{4.13}$$

This is just a statement about conservation of the droplet volume $V = \int h_0\,dx = \pi R^3/4$ since

$$V_t = -\int_{\partial D} J_q \cdot \mathbf{n}\,ds. \tag{4.14}$$

Inner products of (4.1) with $\psi_{x,y}$ determine the translation dynamics. Integration with ψ_x gives

$$
\begin{aligned}
X_t \cdot \hat{\mathbf{x}} \int_D \psi_x \frac{\partial h_0}{\partial x} d\mathbf{x} &= \int_{\partial D} \psi_x h_0^q \nabla \Delta h_q \cdot \mathbf{n}\, ds - \int_{\partial D} h_0^q (\Delta h_q) \nabla \psi_x \cdot \mathbf{n} ds \\
&\quad - \int_{\partial D} x \nabla h_q \cdot \mathbf{n} ds + \int_{\partial D} h_q \cos\theta \, ds \\
&\equiv \int_{\partial D} \psi_x J_q \cdot \mathbf{n}\, ds + B_1 + B_2 + B_3.
\end{aligned} \tag{4.15}
$$

In writing this, the inner products with $\partial h_0/\partial y$ and $\partial h_0/\partial R$ are zero by symmetry. We will argue that $B_1 = B_2 = B_3 = 0$.

First, since the leading order flux is constant across region II,

$$J_q \cdot \mathbf{n} = h_0^q \nabla \Delta h_q \cdot \mathbf{n} = \mathcal{O}(1), \quad r \to R. \tag{4.16}$$

Therefore

$$\nabla \Delta h_q \cdot \mathbf{n} = \mathcal{O}(|r-R|)^{-q}, \quad \Delta h_q = \mathcal{O}(|r-R|^{-q+1}), \quad r \to R. \tag{4.17}$$

Using (4.10), this means that

$$h_0^q (\psi_x)_z \Delta h_q = \mathcal{O}(|r-R|^{-q+2}), \quad r \to R, \tag{4.18}$$

so that integral $B_1 = 0$.

For the integrals B_2 and B_3, consider first the subcase $q = 1$. The relevant matching conditions are

$$h_1(R,\theta) = \lim_{z\to\infty} H_1(z), \quad (h_1)_r(R,\theta) = \lim_{z\to\infty} H_2'(z). \tag{4.19}$$

Since H_1 and H_2 are independent of θ, the integrals B_2 and B_3 vanish by symmetry.

For noninteger q, the terms in the region II expansion necessary for matching would be orders ϵ^q and ϵ^{q+1}. If such orders were included in the expansion, they would solve equations like

$$\Big(H_1^q [(H_n)_{zz} - U''(H_1) H_n]_z \Big)_z = 0, \tag{4.20}$$

where $1 < n < 3$. Since there is no flux of order $\epsilon^{q+n-3} > \epsilon^q$, integrating (4.20), one gets

$$(H_n)_{zz} - U''(H_1) H_n = \text{constant}. \tag{4.21}$$

From this it is seen that the solution H_n of *any* order $n < 3$ is independent of θ, and therefore the integrals B_2 and B_3 again vanish.

We now return to determining the migration dynamics. A similar argument as presented holds for the inner product with ψ_y. Using (4.11), (4.15) and $\mathbf{n} = (\cos\theta, \sin\theta)$ leads to

$$X_t = -\frac{1}{\pi} \frac{R\Psi(R)}{\int_0^R \Psi(r) r^2 dr} \left(\int_{\partial D} \mathbf{n} J_q \cdot \mathbf{n}\, ds \right). \tag{4.22}$$

Equations (4.13), (4.22) specify the droplet dynamics once the boundary value problem (3.16)–(3.18) is solved.

4.2. Case $q \in (2, 3)$. As for the case of $p \le 2$, in region II the flux of order ϵ^q involves the correction to H of order ϵ^3, which satisfies the linear equation

$$0 = (J_q \cdot \hat{\mathbf{z}})_z = \Big(H_1^q \Big[(H_3)_{zz} - R^{-1}(H_2)_z - zR^{-2}(H_1)_z + \frac{1}{2}U'''(H_1)H_2^2 + U''(H_1)H_3 \Big]_z \Big)_z . \tag{4.23}$$

The relevant solvability condition for the linear equation (4.23) is found by using the bounded function

$$\Phi(z) = -\int_z^\infty \frac{H_1 - 1}{H_1^q} dz', \tag{4.24}$$

which is in the adjoint nullspace of the linear operator in (4.23) and corresponds to translation. Taking an inner product with (4.23) gives

(4.25)
$$0 = \Big[(J_q \cdot \hat{\mathbf{z}})\, \Phi - (H_1 - 1)[(H_3)_{zz} - U''(H_1)H_3] + (H_1)_z(H_3)_z - (H_1)_{zz}(H_3) \Big]_{-\infty}^{\infty} + Q,$$
$$Q = \int_{-\infty}^{\infty} (H_1 - 1) \Big[R^{-1}(H_2)_z + zR^{-2}(H_1)_z + \frac{1}{2}U'''(H_1)H_2^2 \Big]_z dz.$$

Here $(J_q \cdot \hat{\mathbf{z}}) = -(J_q \cdot \mathbf{n})$ is just the flux matched to the region III solution at $z = -\infty$. Note that the term Q inherits radial symmetry from H_1, H_2 and therefore should be inconsequential for migration dynamics.

Applying the far field and matching conditions

$$\Phi \sim 0, \quad (H_1)_z \sim 1, \quad (H_3)_z \sim +\frac{\partial^2 h_1}{\partial r^2}(R,\theta)z - \frac{\partial h_2}{\partial r}(R,\theta), \quad z \to \infty, \tag{4.26}$$

$$H_1 \sim 1, \quad z \to -\infty, \tag{4.27}$$

to (4.25) gives

$$\Big(J_q \cdot \hat{\mathbf{z}} \Big) \int_{-\infty}^{\infty} \frac{H_1 - 1}{H_1^q} dz' = \frac{\partial h_2}{\partial r}(R, \theta) - Q. \tag{4.28}$$

In region I, we expand $h = h_0(x,t) + \epsilon h_1 + \epsilon^2 h_2 + o(\epsilon^2)$ and obtain the same as (4.1), except that it applies to the correction at order ϵ^2 instead of order ϵ^q:

$$\mathcal{L}h_2 = X_t \cdot \nabla h_0 - R_t \frac{\partial h_0}{\partial R}, \quad \mathcal{L} = \nabla\cdot[h_0^q \nabla\Delta, \quad \mathbf{x} \in D. \tag{4.29}$$

Solvability conditions are obtained as in case $q < 2$. An inner product with ψ_R gives

$$R_t = -\frac{\int_{\partial D} h_0^q \nabla\Delta h_2 \cdot \mathbf{n}\, ds}{\int_D \partial h_0/\partial R \, dx}. \tag{4.30}$$

Note that $h_0^q \nabla\Delta h_2$ would be the flux at order ϵ^2, but this is zero since the leading order flux scales like ϵ^q. This means that $R_t = 0$ on the timescale specified by $\tau(\epsilon)$. One could potentially obtain the slow dynamics for mass exchange by going further in the expansion, where a result like (4.13) should follow on a timescale ϵ^q instead of $\tau(\epsilon)$.

For exponents $q \geq 2$ the functions ψ_x, ψ_y are not bounded at $r = R$, but we can integrate over a smaller disk D_ρ of radius ρ and take $\rho \to R$. To avoid excessive notation, the integrals $\int_D, \int_{\partial D}$ which appear below should be interpreted as this limit. Integration with ψ_x yields

$$\begin{aligned} X_t \cdot \hat{\mathbf{x}} \int_D \psi_x (h_0)_x \, d\mathbf{x} &= \int_{\partial D} \psi_x h_0^q \nabla \Delta h_2 \cdot \mathbf{n} ds - \int_{\partial D} h_0^q (\Delta h_2) \nabla \psi_x \cdot \mathbf{n} ds \\ &\quad - \int_{\partial D} x \nabla h_2 \cdot \mathbf{n} ds + \int_{\partial D} h_2 \cos\theta \, ds \\ &\equiv B_1 + B_2 + B_3 + B_4. \end{aligned} \tag{4.31}$$

In contrast to $p < 2$, only the boundary term B_3 is not zero. In light of (4.9), one has the asymptotics

$$\psi_x = \mathcal{O}(|r-R|^{2-q}), \quad \nabla \psi_x \cdot \mathbf{n} = \mathcal{O}(|r-R|^{1-q}), \quad h_0^q = \mathcal{O}(|r-R|^q). \tag{4.32}$$

The boundedness of derivatives of h_2 then implies $B_1 = B_2 = 0$. For B_4, matching to region I implies $H_2 \sim \frac{1}{2}(h_0)_{rr}(R,\theta) z^2 - (h_1)_r(R,\theta) z + (h_2)(R,\theta)$ for large z. This means that h_2 is independent of θ, and symmetry gives $B_4 = 0$. It follows that

$$X_t \cdot \hat{\mathbf{x}} \int_D \psi_x (h_0)_x \, d\mathbf{x} = - \int_{\partial D} x \nabla h_2 \cdot \mathbf{n} ds. \tag{4.33}$$

A similar expression can be obtained using ϕ_y. Combining this with (4.28), the terms involving Q drop away by symmetry, leaving

$$X_t = -\frac{1}{\pi} \frac{R^2 \int_{-\infty}^{\infty} H_1^{1-q} - H_1^{-q} dz'}{\int_0^R \Psi(r) r^2 dr} \left(\int_{\partial D} \mathbf{n} J_q \cdot \mathbf{n} ds \right). \tag{4.34}$$

4.3. Case $q > 3$. The expansion in region II is now done as $H = \epsilon H_1 + \epsilon^2 H_2 + \epsilon^3 H^3 + \cdots$. At the level of the first nonzero flux J_q, we get the linear equation

$$\begin{aligned} -(X_t \cdot \mathbf{n})(H_1)_z = \Bigg(H_1^q \Bigg[& (H_3)_{zz} - R^{-1}(H_2)_z - z R^{-2} (H_1)_z \\ & + \frac{1}{2} U'''(H_1) H_2^2 + U''(H_1) H_3 \Bigg]_z \Bigg)_z . \end{aligned} \tag{4.35}$$

The solvability argument proceeds as for the case $2 < q < 3$ and uses the bounded function Φ defined in (4.24). The inner product with (4.35) gives the same result as for $2 < q < 3$ except that the left-hand side is nonzero:

$$(X_t \cdot \mathbf{n}) \int_{-\infty}^{\infty} \frac{H_1 - 1}{H_1^{q-1}} dz = -(J_q \cdot \mathbf{n}) \int_{-\infty}^{\infty} \frac{H_1 - 1}{H_1^q} dz - \frac{\partial h_2}{\partial r}(R,\theta) - Q. \tag{4.36}$$

Here $(J_q \cdot \mathbf{n}) = -(J_q \cdot \hat{\mathbf{z}})(z = -\infty)$ is the flux matched to region III.

The expansion in region I is $h = h_0 + \epsilon h_1 + \epsilon^2 h_2 + o(\epsilon^2)$, which means h_2 solves

$$\nabla \cdot (h_0^q \nabla \Delta h_2) = 0. \tag{4.37}$$

This is the homogeneous version of (4.29), and therefore the relevant solvability conditions (for each coordinate direction) are the same as (4.33) with the left-hand side suppressed:

$$\int_{\partial D} x \nabla h_2 \cdot \mathbf{n} ds = 0 = \int_{\partial D} y \nabla h_2 \cdot \mathbf{n} ds. \tag{4.38}$$

We can now multiply (4.36) by x or y and integrate over ∂D and use (4.38). Again the Q term drops away and we are left with

$$X_t = -\frac{1}{\pi}\frac{\int_{-\infty}^{\infty}[H_1^{1-q} - H_1^{-q}]dz'}{R\int_{-\infty}^{\infty}[H_1^{2-q} - H_1^{1-q}]dz'}\left(\int_{\partial D} \mathbf{n}J_q \cdot \mathbf{n} ds\right). \tag{4.39}$$

4.4. Case $q = 2$. This case is similar to $q \in (2, 3)$, but there are logarithmically diverging terms that require care. The flux of order ϵ^2 in region II satisfies the linear equation

$$\begin{aligned} 0 = (J_2 \cdot \hat{\mathbf{z}})_z = \Bigg(H_1^2 \Big[&(H_3)_{zz} - R^{-1}(H_2)_z - zR^{-2}(H_1)_z \\ &+ \frac{1}{2}U'''(H_1)H_2^2 + U''(H_1)H_3\Big]_z\Bigg)_z, \end{aligned} \tag{4.40}$$

which again says that the normal component of the flux is constant. The relevant solvability condition uses the function

$$\Phi = \int_{-\infty}^{z} \frac{H_1 - 1}{H_1^2} dz', \tag{4.41}$$

which diverges logarithmically:

$$\Phi = \ln(z) + O(1), \quad z \to \infty. \tag{4.42}$$

Multiplying Φ by (4.40) and integrating from $-\infty$ to some finite value $z = Z$ (since the result is unbounded as $Z \to \infty$) gives a result similar to (4.25):

$$\begin{aligned} &0 = \Big[(J_2 \cdot \hat{\mathbf{z}})\,\Phi - (H_1 - 1)[(H_3)_{zz} - U''(H_1)H_3] + (H_1)_z(H_3)_z - (H_1)_{zz}(H_3)\Big]_{-\infty}^{Z} + Q, \\ &Q = \int_{-\infty}^{Z} (H_1 - 1)\left[R^{-1}(H_2)_z + zR^{-2}(H_1)_z + \frac{1}{2}U'''(H_1)H_2^2\right]_z dz. \end{aligned} \tag{4.43}$$

Since the flux J_2 is nonzero, integrating (4.40) directly gives $(H_3)_{zzz} \sim 1/z^2$ for large z. Therefore H_3 is bounded and $(H_3)_z$ diverges logarithmically as $z \to +\infty$. The balance of logarithmically diverging terms in (4.43) gives

$$(H_3)_z = (J_2 \cdot \mathbf{n}) \ln(z) + O(1), \quad z \to \infty, \tag{4.44}$$

where $(J_q \cdot \mathbf{n}) = -(J_q \cdot \hat{\mathbf{z}})(z = -\infty)$ is the flux matched to region III.

In region I, we expand $h = h_0(x, t) + \epsilon h_1 + \epsilon^2 \ln(1/\epsilon)h_* + o(\epsilon^2 \ln(1/\epsilon))$. Then h_* solves the linear equation (4.1), and the solvability arguments proceed as before. Like the case $q \in (2, 3)$, $R_t = 0$ to leading order (albeit mass exchange is only logarithmically slower). The other solvability conditions are obtained by taking inner products with ψ_x, ψ_y, which produces a result analogous to (4.33):

$$X_t \cdot \hat{\mathbf{x}} \int_D \psi_x(h_0)_x \, d\mathbf{x} = -\int_{\partial D} x\nabla h_* \cdot \mathbf{n} ds. \tag{4.45}$$

Matching conditions that relate $(h_*)_x$ to $(H_3)_z$ are now derived. It is assumed that region I and II solutions describe the same solution on some overlapping region

$1 \ll z \ll [\epsilon \ln(1/\epsilon)]^{-1}$. Within this region, a Taylor expansion is justified for h_0, h_1 but not h_* so that

$$\begin{aligned}(4.46)\qquad &(H_1)_z + \epsilon(H_2)_z + \epsilon^2(H_3)_z + o(\epsilon^2)\\ &= -(h_0)_r - \epsilon(h_1)_r - \epsilon^2\ln(1/\epsilon)(h_*)_r + o(\epsilon^2\ln(1/\epsilon))\\ &= -(h_0)_r(R,\theta) - \epsilon\Big[(h_1)_r(R,\theta) + (h_0)_{rr}(R,\theta)z\Big]\\ &\quad - \epsilon^2\ln(1/\epsilon)(h_*)_r(R,\theta) + o(\epsilon^2\ln(1/\epsilon)).\end{aligned}$$

Equating terms at order 1 and ϵ gives the usual matching conditions for regular expansions. For the logarithmic terms, the procedure is to take $\epsilon \to 0$ and $z \sim [\epsilon\ln(1/\epsilon)]^{-1}$ simultaneously. Using (4.44), for large z we have

$$(4.47)\quad (H_3)_z = (J_2\cdot\mathbf{n})\ln\Big([\epsilon\ln(1/\epsilon)]^{-1}\Big) + O(1) = (J_2\cdot\mathbf{n})\ln(1/\epsilon) + \mathcal{O}\Big(\ln(\ln(1/\epsilon))\Big).$$

Inserting into (4.46) and equating terms of order $\epsilon^2\ln(1/\epsilon)$ gives

$$(4.48)\qquad (J_2\cdot\mathbf{n}) = -(h_*)_r(R,\theta).$$

This can be combined with (4.45) to yield

$$(4.49)\qquad X_t = -\frac{1}{\pi}\frac{R^2}{\int_0^R \Psi(r)r^2dr}\left(\int_{\partial D}\mathbf{n}J_q\cdot\mathbf{n}\,ds\right).$$

4.5. Case $q = 3$. This case is similar to both $q > 3$ and $q \in (2,3)$, but there are again logarithmically diverging terms. The flux of order ϵ^3 in region II satisfies the linear equation

$$\begin{aligned}0 = (J_3\cdot\hat{\mathbf{z}})_z = \Bigg(H_1^3\Bigg[&(H_3)_{zz} - R^{-1}(H_2)_z - zR^{-2}(H_1)_z\\ (4.50)\qquad &+ \frac{1}{2}U'''(H_1)H_2^2 + U''(H_1)H_3\Bigg]_z\Bigg)_z.\end{aligned}$$

A solvability argument identical to the case $q \in (2,3)$ produces

$$(4.51)\qquad (J_3\cdot\mathbf{n})\int_{-\infty}^{\infty}\frac{H_1-1}{H_1^3}dz' = (H_3)_z - Q.$$

In region I, we expand $h = h_0(x,t)+\epsilon h_1+\epsilon^2/\ln(1/\epsilon)h_*+o(\epsilon^2/\ln(1/\epsilon))$, so that h_* solves the linear equation (4.1) with $q = 3$, and the solvability arguments proceed as before. As for all cases $q \geq 2$, $R_t = 0$ to leading order. In this case, the inner products with ψ_x, ψ_y diverge logarithmically, so we integrate on a disk $D(\rho)$ with radius $\rho < R$ and consider the asymptotics as $\rho \to R$. Multiplying by ψ_x and integrating gives

(4.52)
$$\begin{aligned}(X_t\cdot\hat{\mathbf{x}})\int_{D(\rho)}\psi_x(h_0)_x\,d\mathbf{x} &= \int_{\partial D(\rho)}\psi_x h_0^3\nabla\Delta h_*\cdot\mathbf{n}ds - \int_{\partial D(\rho)}h_0^3(\Delta h_*)\nabla\psi_x\cdot\mathbf{n}ds\\ &- \int_{\partial D(\rho)}x\nabla h_*\cdot\mathbf{n}ds + \int_{\partial D(\rho)}h_*\cos\theta\,ds = B_1+B_2+B_3+B_4.\end{aligned}$$

The integral on the left-hand side has a logarithmic singularity as $\rho \to R$ because of (4.9); in particular,

$$\int_{D(\rho)} \psi_x (h_0)_x \, d\mathbf{x} = -\pi R^2 \ln |R-\rho| + \mathcal{O}(1), \quad \rho \to R. \tag{4.53}$$

Matching conditions (which are detailed below) require $\nabla h_* \sim C \ln |R-r|$. As a consequence, we find that h_* is bounded and

$$\nabla \Delta h_* \sim |R-r|^{-2}, \quad \Delta h_* \sim |R-r|^{-1} \tag{4.54}$$

as $r \to R$. All this implies that the integrals B_1, B_2, B_4 are bounded as $\rho \to R$ but B_3 diverges logarithmically. Using (4.52)–(4.53) gives

$$\int_{\partial D(\rho)} x \nabla h_* \cdot \mathbf{n} ds = \pi R^2 (X_t \cdot \hat{\mathbf{x}}) \ln |R-\rho| + \mathcal{O}(1), \quad \rho \to R. \tag{4.55}$$

The matching condition that relates $(h_*)_r$ to $(H_3)_z$ is derived as for $q = 2$. Equating expansions for h_r in regions I and II, then for $1 \ll z \ll \log(1/\epsilon)$,

$$\begin{aligned} &(H_1)_z + \epsilon (H_2)_z + \epsilon^2 (H_3)_z + o(\epsilon^2) \\ &= -(h_0)_r - \epsilon (h_1)_r - \epsilon^2 / \ln(1/\epsilon) (h_*)_r + o(\epsilon^2/\ln(1/\epsilon)) \\ &= -(h_0)_r (R,\theta) - \epsilon \Big[(h_1)_r (R,\theta) + (h_0)_{rr} (R,\theta) z \Big] \\ &\quad - \epsilon^2 (h_1)_{rr} (R,\theta) z - \epsilon^2/\ln(1/\epsilon)(h_*)_r + o(\epsilon^2/\ln(1/\epsilon)). \end{aligned} \tag{4.56}$$

Let $(h_*)_r \sim C \ln |R-r|$, $r \to R$, where C is to be determined. Taking $\epsilon \to 0$ with $z \sim \ln(1/\epsilon)$ simultaneously implies for large z

$$(h_*)_r = C \ln(\epsilon z) + \mathcal{O}(1) = C \ln(\epsilon) + \mathcal{O}(\ln(\ln(1/\epsilon))), \quad \epsilon \to 0. \tag{4.57}$$

Inserting into (4.56) and equating terms of order ϵ^2, we obtain

$$C = \lim_{z \to \infty} -(H_3)_z. \tag{4.58}$$

Finally, combining (4.51), (4.55), (4.58),

$$X_t = -\frac{1}{\pi} \frac{\int_{-\infty}^{\infty} [H_1^{-2} - H_1^{-3}] dz}{R} \left(\int_{\partial D} \mathbf{n} J_q \cdot \mathbf{n} \, ds \right). \tag{4.59}$$

5. Effective medium approximation and identification of timescales.

One potentially useful approximation to the free boundary problem described in section 2 utilizes Green's functions similar to the effective medium approximations for standard Ostwald ripening [38]. This is employed to determine timescales and study large systems of interacting droplets.

5.1. Reduced system.

Let X_k, R_k, $k = 1, \ldots, N$, be the droplet centers and radii. We want to solve $\Delta P = 0$ exterior to the droplets, i.e., for all $\mathbf{x}$, $|\mathbf{x} - X_k| > R_k$, subject to the boundary conditions

$$P(\mathbf{x}) = \frac{2}{R_k}, \quad |\mathbf{x} - X_k| = R_k. \tag{5.1}$$

The simplest approximation looks for a solution as a sum of Green's functions,

$$P(\mathbf{x}) = B_0 + \sum_{k=1}^{N} B_k \ln|\mathbf{x} - X_k|^2. \tag{5.2}$$

For each $j = 1, \dots, N$, the boundary condition which one wishes to satisfy is

$$\frac{2}{R_j} = B_0 + \sum_{k=1}^{N} B_k \ln|\mathbf{x} - X_k|^2 \quad \text{for} \;\; |\mathbf{x} - X_j| = R_j. \tag{5.3}$$

Assuming the droplets are well separated, the approximation $|\mathbf{x} - X_k| \approx |X_j - X_k|$ holds on the boundary of droplet $j \neq k$, giving

$$\frac{2}{R_j} = B_0 + B_j \ln(R_j^2) + \sum_{k \neq j} B_k \ln|X_j - X_k|^2, \quad j = 1, \dots, N. \tag{5.4}$$

The system is completed by the requirement that there be no flux at infinity:

$$\int_S \nabla P \cdot n \to 0, \tag{5.5}$$

as the curve S (take it to be a giant circle) is taken out to infinity. As $\mathbf{x} \to \infty$, $1/|\mathbf{x} - R_k| \approx 1/|\mathbf{x}|$, and therefore

$$\int_S \nabla P \cdot n \to \sum_{k=1}^{N} B_k \left(\int_S \frac{1}{|\mathbf{x}|} ds \right) = 2\pi \sum_{k=1}^{N} B_k. \tag{5.6}$$

This integral will be zero only if

$$\sum_{k=1}^{N} B_k = 0. \tag{5.7}$$

Equations (5.4) and (5.7) define an $(N+1) \times (N+1)$ linear problem to be solved.

The integral in (2.5) to be computed for each j is

$$\int_{\partial D_j} J \cdot \mathbf{n} \, ds = - \int_{\partial D_j} \nabla P \cdot \mathbf{n} \, ds = 4\pi B_j. \tag{5.8}$$

The integral in (2.6) to be evaluated for each j is

$$\begin{aligned} \int_{\partial D_j} (J \cdot \mathbf{n})\mathbf{n} \, ds &= - \int_{\partial D_j} \left(\sum_k B_k \frac{2(\mathbf{x} - X_k) \cdot \mathbf{n}}{|\mathbf{x} - X_k|^2} \right) ds(\mathbf{x}) \\ &\approx -2 \left(\sum_{k \neq j} B_k \frac{X_j - X_k}{|X_j - X_k|^2} \right) \cdot \left(\int_{\partial D_j} \mathbf{n} \otimes \mathbf{n} \, ds \right), \end{aligned} \tag{5.9}$$

where the same approximation $|\mathbf{x} - X_k| \approx |X_j - X_k|$ as before was used. Since

$$\int_{\partial D_j} \mathbf{n} \otimes \mathbf{n} \, ds = \pi R_j \mathbf{I}, \tag{5.10}$$

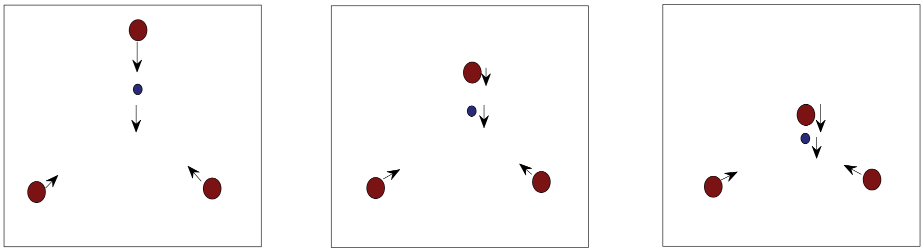


FIG. 5.1. *Example numerical solution of an approximate system with four initial droplets. The net repulsion experienced by the drop in the center is small, allowing one of the large droplets to catch up.*

where $\mathbf{I}$ is the identity matrix, it follows that

$$\int_{\partial D_j} (J \cdot \mathbf{n})\mathbf{n}\, ds \approx -2\pi R_j \sum_{k \neq j} B_k \frac{X_j - X_k}{|X_j - X_k|^2}. \tag{5.11}$$

It is instructive to examine a simple situation where only two droplets interact. By virtue of the quasi-steady diffusion (2.4) it follows that the flux J will on average be toward the smaller drop. In the context of the above approximation, that means that if $R_1 > R_2$, then $B_1 > 0 > B_2$. Using (2.6) and (5.11), it follows that the velocity of both droplets is in the direction of the *smaller* droplet.

By virtue of the mobility factor (2.7), a smaller droplet moves faster, and therefore it would simply "run away" from a single large droplet. On the other hand, if a smaller droplet is surrounded by several large droplets, the net repulsion can be small enough so that merging with a larger droplet is possible. Figure 5.1 shows a numerically computed example with three large droplets and a small droplet, which eventually touches one of the larger droplets. Simulations of the thin film equation indicate that this initiates a rapid coalescence of both drops [29].

5.2. Dynamic timescales. Consider now a reasonably large array of droplets which all have a similar size R and typical spacing L, so that the volume per unit area is

$$H_{average} = \frac{R^3}{L^2}, \tag{5.12}$$

which is constant as time progresses.

Timescale for ripening. The approximation (5.2) gives the scaling

$$B_j \sim R^{-1}/\ln L, \tag{5.13}$$

which with (5.8) further implies

$$\int_{\partial D_j} J \cdot \mathbf{n}\, ds \sim R^{-1}/\ln L. \tag{5.14}$$

For exponents $q < 2$, using (2.5), the timescale for ripening (i.e., mass exchange) can be computed as

$$\tau_{ripe} \sim \frac{R}{R_t} \sim R^4 \ln L \sim H_{average}^{4/3} L^{8/3} \ln L, \quad q < 2. \tag{5.15}$$

For exponents $q \geq 2$, the ripening dynamics occur on a timescale of the flux, i.e., ϵ^q rather than $\tau(\epsilon)$. This can be accommodated by including an extra factor in the timescale:

$$\tau_{ripe} \sim \frac{R}{R_t} \sim \frac{\tau(\epsilon)}{\epsilon^q} H_{average}^{4/3} L^{8/3} \ln L, \quad q \geq 2. \tag{5.16}$$

Timescale for migration. Using (5.11) and (5.13), one can obtain

$$\int_{\partial D_j} (J \cdot \mathbf{n})\mathbf{n}\, ds \sim L^{-1}/\ln L. \tag{5.17}$$

Using (2.6), the timescale for migration can be computed as

$$\tau_{mig} \sim \frac{L}{X_t} \sim \begin{cases} H_{average}^{2/3} L^{10/3} \ln L, & q < 2, \\ H_{average}^{\frac{4-q}{3}} L^{(14-2q)/3} \ln L, & q \in [2,3], \\ H_{average}^{1/3} L^{8/3} \ln L, & q > 3. \end{cases} \tag{5.18}$$

The limit of large droplet size in the unscaled equation. Here we show that our scaling results are, suitably interpreted, the same as those derived in the companion paper [9]. The starting point there was the unscaled thin film equation

$$h_t = \nabla \cdot (h^q \nabla p), \quad p = U'(h) - \Delta h, \quad q > 0. \tag{5.19}$$

In [9], the limit of large droplet volume was considered, in contrast to a small precursor film. In this case, let $V \gg 1$ be a typical droplet volume with characteristic interdroplet distance L'. This suggests that the natural small parameter is $\epsilon = V^{-1/3}$. Rescaling (5.19) using

$$x \to \epsilon^{-1} x, \quad t \to \tau(\epsilon)^{-1} \epsilon^{q-4}, \quad h \to h\epsilon^{-1} \tag{5.20}$$

gives exactly (1.1). The average droplet size after rescaling is $R = 1$, and the characteristic distance between droplets is

$$L = \epsilon L' = \frac{V^{1/6}}{\overline{H}^{1/2}}, \tag{5.21}$$

where $\overline{H} = V/(L')^2$. The mass density for the scaled equation is

$$H_{average} = \frac{1}{\epsilon^2 (L')^2} = \frac{\overline{H}}{V^{1/3}}. \tag{5.22}$$

Timescales with respect to the unscaled equation (5.19) can now be written in terms of V and $\overline{H}$. For the ripening times given by either (5.15) or (5.16) one obtains

$$\tau_{ripe}^{unscaled} = \tau(\epsilon)\epsilon^{4-q} \tau_{ripe} \sim V^{4/3} \ln V + \mathcal{O}(1), \quad V \to \infty. \tag{5.23}$$

For the migration timescale (5.18) one has

$$\tau_{mig}^{unscaled} = \tau(\epsilon)\epsilon^{4-q} \tau_{mig} = \mathcal{O}(1) + \frac{1}{\overline{H}} \begin{cases} V^{5/3} \ln V, & q \in (0,2), \\ V^{5/3}, & q = 2, \\ V^{\frac{7-q}{3}} \ln V, & q \in (2,3), \\ V^{4/3} \ln^2 V, & q = 3, \\ V^{4/3} \ln V, & q > 3. \end{cases} \quad V \to \infty, \tag{5.24}$$

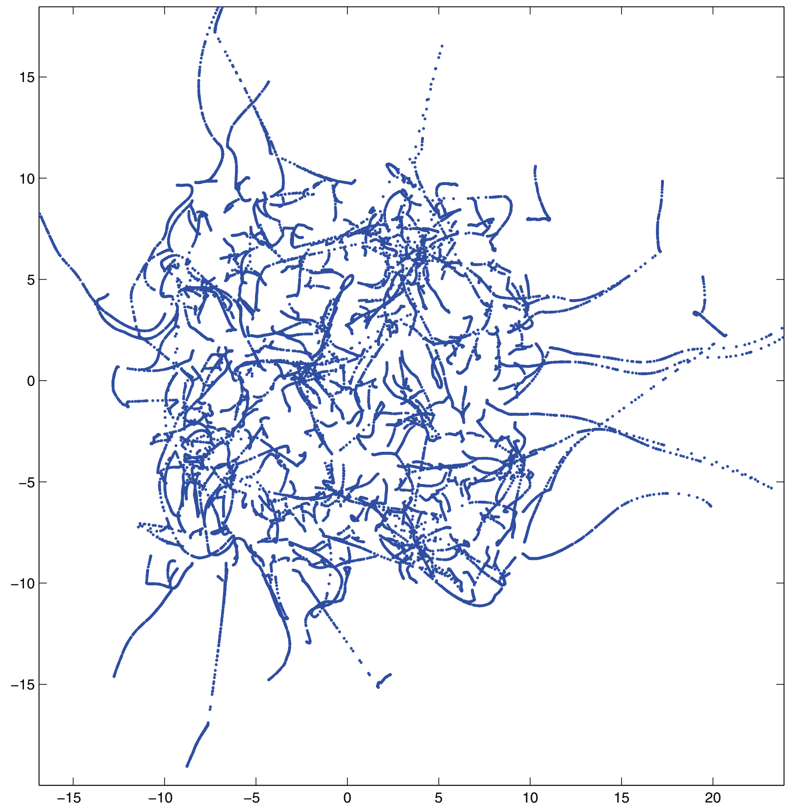


Fig. 6.1. *Trajectories for a simulation with* 500 *initial droplets* ($q = 3$)*. The direction of motion is mostly outward from the center. Note that small, uncoalesced droplets on the fringes are repelled.*

6. Large scale coarsening by coalescence. We conclude by using the approximations of the previous section to study the evolution of a large assembly of droplets. We focus on the most relevant mobility exponent $q = 3$, which corresponds to a fluid with Newtonian viscosity and a no-slip boundary condition. In doing this, the exchange of material between droplets is ignored, and only the leading order effect, migration, is considered. There are no boundaries in the calculation, so that (5.5) applies. An ad hoc criterion for coalescence is applied, which states that when the perimeters of two droplets overlap, their volume is combined and the center of mass is preserved.

Figure 6.1 is an illustration of the dynamics. The simulation was started with 500 droplets in random locations, each with a random but nearly uniform radius. Droplets in the middle of the assembly coalesce first, simply because they have a greater number of neighbors. As time progresses, it follows that smaller, more mobile droplets on the fringes will be driven away, since the motion is opposite the flux, which is toward larger droplets. The amount of time that droplets take to move (relative to the interdroplet distance) increases since the driving force given by flux decreases with increasing droplet size.

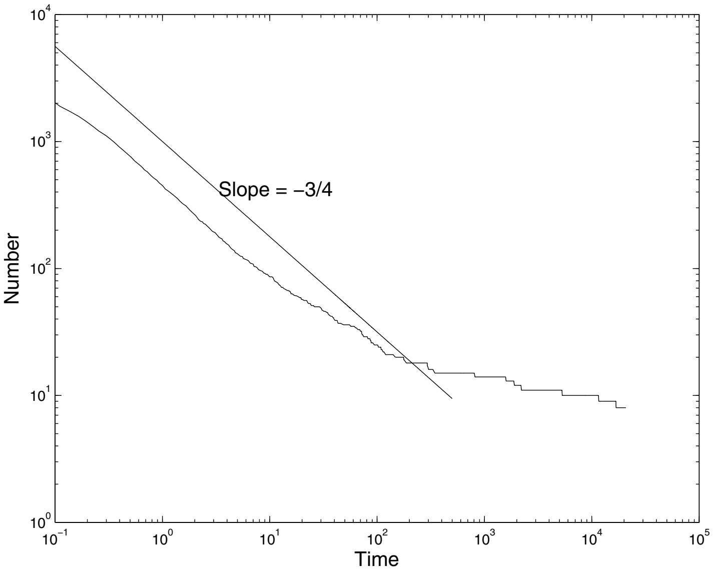


FIG. 6.2. *Dynamic scaling of coalescence-driven coarsening ($q = 3$), using* 5000 *initial droplets. A line with slope* $-3/4$ *is provided for comparison to the predictions of* [29].

Figure 6.2 shows the droplet number plotted as a function of time, for a simulation with 5000 droplets initially. Dynamic scaling of the coarsening process was predicted [29]; in particular, the relevant length scale (the typical interdroplet distance L) should increase as $t^{3/8}$. Since the number of droplets N scales according to

$$NL^2 \sim \text{area of domain},$$

N should scale in time like $t^{-3/4}$. This is more or less borne out by the results in Figure 6.2. At late times when the array has spread out to a somewhat larger area, there is a slowing of the coarsening process, also seen in the computational results.

7. Conclusions.

The main outcome of this paper is to establish a concrete link between a class of thin film equations and a free boundary problem for the motion of the contact line interface. In contrast to the classical situation of Ostwald ripening, we have shown that migration of droplets, and ultimately coalescence, is a likely mechanism for coarsening. Another feature which distinguishes this problem is its mixed dimensionality: droplets are three-dimensional, but the quasi-steady diffusion of material between them is effectively two-dimensional. This leads to dynamic scaling with exponents different than the familiar "1/3" power law.

There is some experimental support for the conclusions which we reach. Limary and Green [15, 16] examined the late stage structural evolution of droplets and measured their size and shape distributions. They found that droplet size (measured as a length scale) evolved as a power law with an exponent that varied from 1/10 to about 2/5. They conclude from their observations that "coarsening occurs via a self-similar, dynamic coalescence process, not Ostwald ripening" (referring to the exchange of material).

As mentioned in the introduction, [32] derives dynamic equations for droplet radius and position in the same thin film model as ours for mobility exponent $q = 3$. The procedure used there does not involve systematic asymptotic expansions and solvability conditions, but rather poses a traveling wave problem that loosely derives from the thin film equation. The authors connect the traveling wave speed to the influence felt by the disjoining pressure at the foot of the droplet. The resulting formulas are vaguely similar to ours but are not quantitatively the same. Most significantly, their assessment of the sign of migration seems to be opposite of ours and contradicts other analytical results [9, 11, 12] as well as direct numerical simulation of the partial differential equation [11, 29]. Indeed, in section IV.B of their paper, they claim that in a two-droplet system "both droplets migrate in the direction of the larger droplet" and show calculations where smaller droplets are attracted to larger droplets. According to (2.6), a larger droplet would move toward a smaller droplet since motion is *opposite* of the flux J. This certainly calls into question the reasoning that leads to their formulation.

There is reason to believe that our results (or at least our analytical procedures) are relevant for a variety of related thin film problems. The main ingredients which we required were the formation of near-equilibrium structures (droplets) and a separation of timescales between their formation and interaction. This separation of timescales is a simple consequence of the nearly degenerate kinetics common to many thin film models. A variety of other phenomena can create instability leading to structure formation and interaction. For example, the Rayleigh–Taylor or Rayleigh–Plateau instabilities lead to formation of migrating liquid ridges [10, 17]. Migration effects similar to ours have also been reported for droplets subject to Marangoni effects [4], chemically driven droplets [31], and two-layer fluids [18]. More broadly, one might expect that degenerate diffusion in other phase separation processes can lead to alternative mechanisms to coarsening.

Acknowledgments. The author is grateful for discussions with Felix Otto, Tobias Rump, and Dejan Slepčev, and for the hospitality afforded during his visit to the University of Bonn.

REFERENCES

[1] J. Becker, G. Grün, R. Seemann, H. Mantz, K. Jacobs, K. R. Mecke, and R. Blossey, *Complex dewetting scenarios captured by thin-film models*, Nature Materials, 2 (2003), pp. 59–62.

[2] A. L. Bertozzi, G. Grün, and T. P. Witelski, *Dewetting films: Bifurcations and concentrations*, Nonlinearity, 14 (2001), pp. 1569–1592.

[3] M. Bestehorn and K. Neuffer, *Surface patterns of laterally extended thin liquid films in three dimensions*, Phys. Rev. Lett., 87 (2001), paper 046101.

[4] M. Bestehorn, A. Pototsky, and U. Thiele, *3D large scale Marangoni convection in liquid films*, European Phys. J. B Condens. Matter Phys., 33 (2003), pp. 457–467.

[5] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system* I: *Interfacial free energy*, J. Chem. Phys., 28 (1957), pp. 258–267.

[6] P. Constantin, T. F. Dupont, R. E. Goldstein, L. P. Kadanoff, M. J. Shelley, and S.-M. Zhou, *Droplet breakup in a model of the Hele-Shaw cell*, Phys. Rev. E, 47 (1993), pp. 4169–4181.

[7] P. G. de Gennes, *Wetting: Statics and dynamics*, Rev. Mod. Phys., 57 (1985), pp. 827–880.

[8] R. Fetzer, K. Jacobs, A. Münch, B. Wagner, and T. P. Witelski, *New slip regimes and the shape of dewetting thin liquid films*, Phys. Rev. Lett., 95 (2005), paper 127801.

[9] K. Glasner, F. Otto, T. Rump, and D. Slepčev, *Ostwald ripening of droplets: The role of migration*, European J. Appl. Math., (2008), to appear.

[10] K. B. Glasner, *The dynamics of pendant droplets on a one-dimensional surface*, Phys. Fluids, 19 (2007), paper 102104.

[11] K. B. Glasner and T. P. Witelski, *Coarsening dynamics of dewetting films*, Phys. Rev. E, 67 (2003), paper 016302.

[12] K. B. Glasner and T. P. Witelski, *Collision versus collapse of droplets in coarsening of dewetting thin films*, Phys. D, 209 (2005), pp. 80–104.

[13] S. Kalliadasis and U. Thiele, eds., *Thin Films of Soft Matter*, Springer, Wien, New York, 2007.

[14] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Chem. Phys. Solids, 19 (1961), pp. 35–50.

[15] R. Limary and P. F. Green, *Late-stage coarsening of an unstable structured liquid film*, Phys. Rev. E, 66 (2002), paper 021601.

[16] R. Limary and P. F. Green, *Dynamics of droplets on the surface of a structured fluid film: Late-stage coarsening*, Langmuir, 19 (2003), pp. 2419–2424.

[17] J. R. Lister, J. M. Rallison, A. A. King, L. J. Cummings, and O. E. Jensen, *Capillary drainage of an annular film: The dynamics of collars and lobes*, J. Fluid Mech., 552 (2006), pp. 311–343.

[18] D. Merkt, A. Pototsky, M. Bestehorn, and U. Thiele, *Long-wave theory of bounded two-layer films with a free liquid-liquid interface: Short- and long-time evolution*, Phys. Fluids, 17 (2005), paper 064104.

[19] V. S. Mitlin, *Dewetting of a solid surface: Analogy with spinodal decomposition*, J. Coll. Int. Sci., 156 (1993), pp. 491–497.

[20] V. S. Mitlin, *Dewetting revisited: New asymptotics of the film stability diagram and the metastable regime of nucleation and growth of dry zones*, J. Coll. Int. Sci., 227 (2000), pp. 371–379.

[21] V. S. Mitlin, *Numerical study of a Lifshitz-Slyozov-like metastable dewetting model*, J. Coll. Int. Sci., 233 (2001), pp. 153–158.

[22] V. S. Mitlin and N. V. Petviashvili, *Nonlinear dynamics of dewetting: Kinetically stable structures*, Phys. Lett. A, 192 (1994), pp. 323–326.

[23] A. Münch, B. Wagner, and T. Witelski, *Lubrication models with small to large slip lengths*, J. Engrg. Math., 53 (2005), pp. 359–383.

[24] B. Niethammer and F. Otto, *Domain coarsening in thin films*, Comm. Pure Appl. Math., 54 (2001), pp. 361–384.

[25] B. Niethammer and R. L. Pego, *On the initial-value problem in the Lifshitz–Slyozov–Wagner theory of Ostwald ripening*, SIAM J. Math. Anal., 31 (2000), pp. 467–485.

[26] L. Onsager, *Reciprocal relations in irreversible processes.* ii., Phys. Rev., 38 (1931), pp. 2265–2279.

[27] A. Oron, *Three-dimensional nonlinear dynamics of thin liquid films*, Phys. Rev. Lett., 85 (2000), pp. 2108–2111.

[28] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Mod. Phys., 69 (1997), pp. 931–980.

[29] F. Otto, T. Rump, and D. Slepčev, *Coarsening rates for a droplet model: Rigorous upper bounds*, SIAM J. Math. Anal., 38 (2006), pp. 503–529.

[30] R. L. Pego, *Front migration in the nonlinear Cahn-Hilliard equation*, Proc. R. Soc. Lond. A, 422 (1989), pp. 261–278.

[31] L. M. Pismen, *Perturbation theory for traveling droplets*, Phys. Rev. E (3), 74 (2006), paper 041605.

[32] L. M. Pismen and Y. Pomeau, *Mobility and interactions of weakly nonwetting droplets*, Phys. Fluids, 16 (2004), pp. 2604–2612.

[33] G. Reiter, *Dewetting of thin polymer films*, Phys. Rev. Lett., 68 (1992), pp. 75–78.

[34] R. Seemann, S. Herminghaus, C. Neto, S. Schlagowski, D. Podzimek, R. Konrad, H. Mantz, and K. Jacobs, *Dynamics and structure formation in thin polymer melt films*, J. Phys. Condensed Matter, 17 (2005), pp. 267–290.

[35] R. A. Segalman and P. F. Green, *Dynamics of rims and the onset of spinodal dewetting at liquid/ liquid interfaces*, Macromolecules, 32 (1999), pp. 801–807.

[36] A. Sharma and R. Khanna, *Pattern formation in unstable thin liquid films*, Phys. Rev. Lett., 81 (1998), pp. 3463–3466.

[37] U. Thiele, M. G. Velarde, and K. Neuffer, *Dewetting: Film rupture by nucleation in the spinodal regime*, Phys. Rev. Lett., 87 (2001), paper 016104.

[38] P. W. Voorhees, *The theory of Ostwald ripening*, J. Statist. Phys., 38 (1985), pp. 231–252.

[39] C. Wagner, *Theorie for alterung von niederschlagen durch umlosen*, Z. Elektrochemie, 65 (1961), pp. 581–594.

[40] S. J. Watson, F. Otto, B. Y. Rubinstein, and S. H. Davis, *Coarsening dynamics of the convective Cahn-Hilliard equation*, Phys. D, 178 (2003), pp. 127–148.

[41] D. E. Weidner and L. W. Schwartz, *Contact-line motion of shear-thinning liquids*, Phys. Fluids, 6 (1994), pp. 3535–3538.

[42] R. Xie, A. Karim, J. F. Douglas, C. C. Han, and R. A. Weiss, *Spinodal dewetting of thin polymer films*, Phys. Rev. Lett., 81 (1998), pp. 1251–1254.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 494–513

NONLINEAR ELECTRON AND SPIN TRANSPORT IN SEMICONDUCTOR SUPERLATTICES*

L. L. BONILLA[†], L. BARLETTI[‡], AND M. ALVARO[†]

Abstract. Nonlinear charge transport in strongly coupled semiconductor superlattices is described by Wigner–Poisson kinetic equations involving one or two minibands. Electron-electron collisions are treated within the Hartree approximation, whereas other inelastic collisions are described by a modified BGK (Bhatnaghar–Gross–Krook) model. The hyperbolic limit is such that the collision frequencies are of the same order as the Bloch frequencies due to the electric field, and the corresponding terms in the kinetic equation are dominant. In this limit, spatially nonlocal drift-diffusion balance equations for the miniband populations and the electric field are derived by means of the Chapman–Enskog perturbation technique. For a lateral superlattice with spin-orbit interaction, electrons with spin up or down have different energies, and their corresponding drift-diffusion equations can be used to calculate spin-polarized currents and electron spin polarization. Numerical solutions show stable self-sustained oscillations of the current and the spin polarization through a voltage biased lateral superlattice thereby providing an example of superlattice spin oscillator.

Key words. quantum drift-diffusion equations, quantum BGK model, Chapman–Enskog method, propagation of pulses, modified Kane model, Rashba spin-orbit interaction, spin oscillator

AMS subject classifications. 34E15, 92C30

DOI. 10.1137/080714312

1. Introduction.

Semiconductor superlattices are essential ingredients in fast nanoscale oscillators, quantum cascade lasers, and infrared detectors. Quantum cascade lasers are used to monitor environmental pollution in gas emissions, to analyze breath in hospitals, and in many other industrial applications [3]. A superlattice (SL) is a convenient approximation to a quasi-one-dimensional (quasi-1D) crystal that was originally proposed by Esaki and Tsu to observe Bloch oscillations, i.e., the periodic coherent motion of electrons in a miniband in the presence of an applied electric field. Figure 1.1(a) shows a simple realization of an N-period SL. Each period of length l consists of two layers of semiconductors with different energy gaps but with similar lattice constants. The SL lengths in the lateral directions, L_y and L_z, are much larger than l, typically tens of microns compared to about ten nanometers. The energy profile of the conduction band of this SL can be modeled as a succession of square quantum wells and barriers along the x direction (Kronig–Penney model) and, for an n-doped SL, we do not have to consider the valence band. A different quasi-1D crystal called a lateral superlattice (LSL) is shown in Figure 1.1(b). In this case, a periodic structure is formed on the top surface of a quantum well (QW), so that L_z is of the order of l and $L_y \gg l$. The wave functions of a single electron in the conduction band

*Received by the editors January 28, 2008; accepted for publication July 21, 2008; published electronically November 14, 2008. This research was supported by the Spanish MECD grant MAT2005-05730-C02-01.
http://www.siam.org/journals/siap/69-2/71431.html

[†]G. Millán Institute of Fluid Dynamics, Nanoscience and Industrial Mathematics, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain (bonilla@ing.uc3m.es, mariano.alvaro@uc3m.es).

[‡]Dipartimento di Matematica "Ulise Dini," Università di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy (barletti@math.unifi.it).

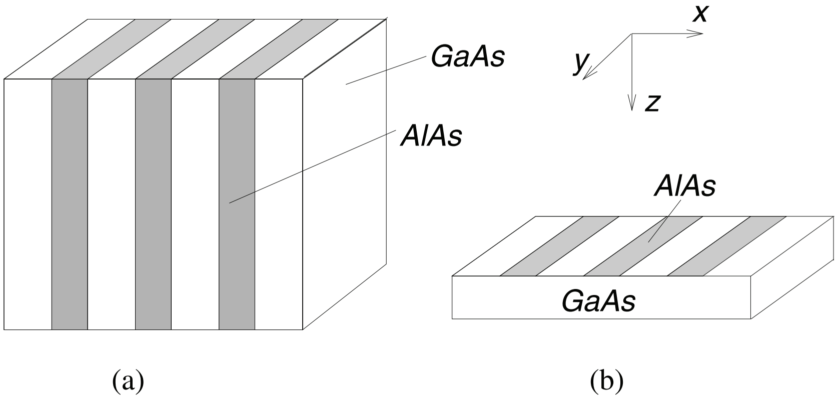


FIG. 1.1. (a) *Schematic drawing of a superlattice.* (b) *A lateral superlattice.*

of an SL can be expanded in terms of 1D Bloch wave functions times plane waves,

$$\frac{1}{\sqrt{S}}\, e^{ik_y y}\psi(z)\, e^{ikx}u_\nu(x,k), \tag{1.1}$$

$$\psi(z) = \begin{cases} e^{ik_z z} & \text{for an SL,} \\ \psi_n(z) & \text{for an LSL,} \end{cases} \tag{1.2}$$

where ν is the miniband index and n is the energy level of the quantum well in the case of an LSL. The function $u_\nu(x,k)$ is l-periodic in x and $2\pi/l$-periodic in k. S is the area of the lateral cross section, equal to L_yL_z for a rectangular cross section.

Many interesting nonlinear phenomena have been observed in voltage-biased SL comprising finitely many periods, including self-oscillations of the current through the SL due to motion of electric field pulses, multistability of stationary charge and field profiles, and so on [3]. It is important to distinguish between strongly and weakly coupled SLs depending on the coupling between their component QWs. Roughly speaking, if barriers are narrow, QWs are strongly coupled and we can use the electronic states (1.1) as a convenient basis in a quantum kinetic description. The resulting reduced balance equations for electron density and electric field are partial differential equations (which may be nonlocal, as we shall see in this paper). On the other hand, for SLs having wide barriers, their QWs are weakly coupled and the electronic states of a single well provide a good basis in a quantum kinetic description, replacing the Bloch functions $e^{ikx}u_\nu(x,k)$ in (1.1). In this case, the balance equations are spatially discrete, and phenomena such as multistability of stationary field profiles, formation and pinning of electric field domains, etc., are theoretically predicted and observed in experiments. See the review [3]. Another promising field of applications is spintronics. Electrons in SLs having at least one period doped with magnetic impurities and subject to a static magnetic field can be distinguished by their spin because the magnetic field splits each miniband in two, giving it different spin-dependent energy [19]. Recently an SL of this type has been proposed as a spin oscillator producing spin-polarized oscillatory currents and able to inject polarized electrons in a contact [6]. Alternatively, materials displaying strong spin-orbit effects can be used as spintronic devices without having to apply magnetic fields; cf. the case of the LSL considered in [14]. In this paper, we will show that an LSL can be used as a spin oscillator.

This paper presents systematic derivations of quantum balance equations for SLs with two populated minibands, and it shows that their numerical solutions may predict space- and time-dependent nonlinear phenomena occurring in these materials. Our methods can be used in 3D crystals, but their application to 1D structures such as SLs and LSLs leads to simpler equations that are less costly to solve. Although nonlinear charge transport in SLs has been widely studied in the last decade (see the reviews [3, 17, 21]), systematic derivations of tractable balance equations for miniband populations and electric field are scarce. One reason is that quantum kinetic equations are nonlocal in space and their collision terms may be nonlocal in space and time [10, 21]. Using them to analyze space- and time-dependent phenomena such as wave propagation or self-sustained oscillations is problematic. In fact, only extremely simple solutions of general quantum kinetic equations (such as thermal equilibrium, disturbances thereof due to weak external fields, and so on) are known; theoretical analysis of these equations is lacking and numerical solutions describing spatiotemporal phenomena are not available. One way to proceed is to adopt simple collision models similar to the Bhatnagar–Gross–Krook (BGK) collision model for classical kinetic theory [1]. We discuss in this paper how to implement a BGK collision model for a quantum kinetic equation that is simple to handle yet keeps an important quantum feature such as the broadening of energy levels [2]. Once we have a quantum kinetic equation for a sufficiently general SL having two minibands, we implement a Chapman–Enskog perturbation procedure to derive the sought balance equations and solve them numerically for realistic SL configurations.

Previous to this work, Lei and coworkers derived quantum hydrodynamic equations describing SL having only one miniband [16, 15]. They use a closure assumption to close a hierarchy of moment equations. For the case of quantum particles in an arbitrary external 3D potential, Degond and Ringhofer [8] have used a similar procedure to derive balance equations. They close the system of moment equations by means of a local equilibrium density obtained by maximizing entropy. The Chapman–Enskog method has been used to derive drift-diffusion equations for single-miniband SLs described by semiclassical [5] and quantum kinetic equations [2]. Earlier, Cercignani, Gamba, and Levermore used the Chapman–Enskog method to derive balance equations for a semiclassical BGK–Poisson kinetic description of a semiconductor with one parabolic band under strong external bias [7].

The rest of this paper is as follows. In section 2, we review the simpler case of nonlinear electron transport in a strongly coupled n-doped SL having only one populated miniband [2]. Starting with a kinetic equation for the Wigner function, we use the Chapman–Enskog perturbation method to derive balance equations for the electron density and the electric field. When these equations are solved numerically for a dc voltage–biased SL with finitely many QWs and realistic parameter values, stable self-sustained oscillations of the current through the SL are found among their solutions, in agreement with experimental observations [2]. Sections 3 to 5 contain the main results of the present work. In section 3, we describe an SL having two populated minibands by proposing a kinetic equation for the Wigner matrix. In section 4, we derive balance equations for the miniband electron populations and the electric field, using an appropriate Chapman–Enskog method and a tight-binding approximation to obtain explicit formulas. The case of an LSL having strong Rashba spin-orbit interaction [18] is important for spintronic applications and has been considered in section 5. We derive and solve numerically the resulting balance equations. Novel self-sustained oscillations of the spin current and polarization are obtained for appropriate values of the parameters. Finally section 6 contains our conclusions, and some technical matters are relegated to the appendix.

2. Single miniband superlattice. The Wigner–Poisson–Bhatnagar–Gross–Krook (WPBGK) system for 1D electron transport in the lowest miniband of a strongly coupled SL is

$$\begin{aligned}(2.1)\quad &\frac{\partial f}{\partial t}+\frac{i}{\hbar}\left[\mathcal{E}\left(k+\frac{1}{2i}\frac{\partial}{\partial x}\right)-\mathcal{E}\left(k-\frac{1}{2i}\frac{\partial}{\partial x}\right)\right]f\\ &+\frac{ie}{\hbar}\left[W\left(x+\frac{1}{2i}\frac{\partial}{\partial k},t\right)-W\left(x-\frac{1}{2i}\frac{\partial}{\partial k},t\right)\right]f\\ &=Q[f]\equiv-\nu_{en}\left(f-f^{FD}\right)-\nu_{\mathrm{imp}}\,\frac{f(x,k,t)-f(x,-k,t)}{2},\end{aligned}$$

$$(2.2)\quad \varepsilon\,\frac{\partial^2 W}{\partial x^2}=\frac{e}{l}\,(n-N_D),$$

$$(2.3)\quad n(x,t)=\frac{l}{2\pi}\int_{-\pi/l}^{\pi/l}f(x,k,t)dk=\frac{l}{2\pi}\int_{-\pi/l}^{\pi/l}f^{FD}(k;n(x,t))dk,$$

$$(2.4)\quad f^{FD}(k;n)=\frac{m^*k_BT}{\pi\hbar^2}\int_{-\infty}^{\infty}\ln\left[1+\exp\left(\frac{\mu-E}{k_BT}\right)\right]\frac{\sqrt{2}\,\Gamma^3/\pi}{[E-\mathcal{E}_1(k)]^4+\Gamma^4}\,dE.$$

Here f, n, N_D, $\mathcal{E}(k)$, d_B, d_W, $l=d_B+d_W$, W, ε, m^*, k_B, T, Γ, ν_{en}, ν_{imp}, and $-e<0$ are the one-particle Wigner function, the 2D electron density, the 2D doping density, the miniband dispersion relation, the barrier width, the well width, the SL period, the electric potential, the SL permittivity, the effective mass of the electron in the lateral directions, the Boltzmann constant, the lattice temperature, the energy broadening of the equilibrium distribution due to collisions [12, p. 28], the frequency of the inelastic collisions responsible for energy relaxation, the frequency of the elastic impurity collisions, and the electron charge, respectively.

The left-hand side of (2.1) can be straightforwardly derived from the Schrödinger–Poisson equation for the wave function in the miniband using the definition of the 1D Wigner function [2]:

$$(2.5)\qquad f(x,k,t)=\frac{2l}{S}\sum_{j=-\infty}^{\infty}\int_{\mathbb{R}^2}\langle\psi^\dagger(x+jl/2,y,z,t)\psi(x-jl/2,y,z,t)\rangle e^{ijkl}d\mathbf{x}_\perp$$

(the second quantized wave function $\psi(x,\mathbf{x}_\perp,t)=\sum_{q,\mathbf{q}_\perp}a(q,q_\perp,t)\phi_q(x)e^{i\mathbf{q}_\perp\cdot\mathbf{x}_\perp}$, $\mathbf{x}_\perp=(y,z)$, is a superposition of the Bloch states corresponding to the miniband and S is the SL cross section [2]). The right-hand side of (2.1) is the sum of $-\nu_e\left(f-f^{FD}\right)$, which represents energy relaxation towards a 1D effective Fermi–Dirac distribution $f^{FD}(k;n)$ (local equilibrium), and $-\nu_i[f(x,k,t)-f(x,-k,t)]/2$, which accounts for impurity elastic collisions [5]. For simplicity, the collision frequencies ν_e and ν_i are fixed constants. Exact and Fermi–Dirac distribution functions have the same electron density, thereby preserving charge continuity as in the classical BGK collision models [1]. The chemical potential μ is a function of n resulting from solving (2.3) with the integral of the collision-broadened 3D Fermi–Dirac distribution over the lateral components of the wave vector $(k,\mathbf{k}_\perp)=(k,k_y,k_z)$:

$$(2.6)\quad f^{FD}(k;n)=\int_{-\infty}^{\infty}\frac{D_\Gamma\left(E-\mathcal{E}_1(k)\right)}{1+\exp\left(\frac{E-\mu}{k_BT}\right)}\,dE,$$

$$(2.7)\quad D_\Gamma(E)=\frac{2}{(2\pi)^2}\int_{\mathbb{R}^2}\delta_\Gamma\left(\frac{\hbar^2\mathbf{k}_\perp^2}{2m^*}-E\right)d\mathbf{k}_\perp=\frac{m^*}{\pi\hbar^2}\int_0^\infty\delta_\Gamma(E_\perp-E)\,dE_\perp.$$

Using the residue theorem for a line-width,

$$\delta_\Gamma(E) = \frac{\sqrt{2}\,\Gamma^3/\pi}{\Gamma^4 + E^4}, \tag{2.8}$$

equation (2.7) yields

$$\begin{aligned} D_\Gamma(E) = \frac{m^*}{\pi\hbar^2}\Bigg\{ 1 + \frac{1}{4\pi}\,\ln\left[\frac{E^2 + \sqrt{2}\Gamma E + \Gamma^2}{E^2 - \sqrt{2}\Gamma E + \Gamma^2}\right] \\ - \frac{\theta(\sqrt{2}|E| - \Gamma)}{2\pi}\left[2\pi - \arctan\left(\frac{\Gamma}{\sqrt{2}|E| + \Gamma}\right) - \arctan\left(\frac{\Gamma}{\sqrt{2}|E| - \Gamma}\right)\right] \\ - \frac{\theta(\Gamma - \sqrt{2}|E|)}{2\pi}\left[\pi + \arctan\left(\frac{\Gamma}{\sqrt{2}E + \Gamma}\right) - \arctan\left(\frac{\Gamma}{\Gamma - \sqrt{2}E}\right)\right] \\ - \frac{\theta(\sqrt{2}E - \Gamma)}{2\pi}\left[\arctan\left(\frac{\Gamma}{\sqrt{2}E + \Gamma}\right) + \arctan\left(\frac{\Gamma}{\sqrt{2}E - \Gamma}\right)\right]\Bigg\}, \end{aligned} \tag{2.9}$$

which is equivalent to (2.4).[1] Here $\theta(E)$ is the Heaviside unit step function. As $\Gamma \to 0+$, the line-width (2.8) tends to the delta function $\delta(E)$, $D_\Gamma(E)$ tends to the 2D density of states, $D(E) = m^*\theta(E)/(\pi\hbar^2)$, and f^{FD} tends to the 3D Fermi–Dirac distribution function integrated over the lateral wave vector $\mathbf{k}_\perp$. In [2], a Lorentzian line-width was used instead of (2.8) and the integral over E in (2.6) extended from 0 to ∞. The integral with the Lorentzian function is not convergent in $E = -\infty$, which is why we prefer using convolution with the "super-Lorentzian" function (2.8) in this work. The integration in (2.7) cannot be carried out explicitly for other standard line-widths such as a Gaussian or a hyperbolic secant. This unnecessarily complicates the numerical integration of the balance equations we will obtain later. Note that, following Ignatov and Shashkin [11], we have not included the effects of the electric potential in our Fermi–Dirac distribution. These model equations can be improved by including scattering processes with change of lateral momentum and an electric field–dependent local equilibrium. However, the resulting model could only be treated numerically and the qualitative features of our derivation and of the nonlocal drift-diffusion equation would be lost in longer formulas.

A different way to introduce a quantum BGK collision model is to define a local equilibrium density matrix operator by minimizing quantum entropy (defined with the opposite sign of the convention that is usual in physics) under constraints giving the electron density and energy density in terms of the density matrix. The resulting expression involves an inverse Wigner transform, and another transform is needed to deduce the local equilibrium Wigner function f^{FD} entering the BGK formula [8]. This f^{FD} is nonlocal in space and can be found only by solving some partial differential equation [8]. As a model for quantum collisions [10, 21], the resulting quantum BGK model is not realistic, in the same way as the original BGK model is not a realistic model for classical collisions. Moreover, the implicit manner in which the model is defined defeats the main asset of the classical BGK collision model: its simplicity, which makes it possible to obtain results analytically. Thus we prefer to introduce a BGK model that can be handled more easily and still incorporates quantum effects. The most important quantum effect affecting the collision term is the broadening of energy levels due to scattering, $\Gamma \approx \hbar/\tau$ (where τ is the lifetime of the level) [12], and

[1] Integrate (2.6) by parts using (2.9).

this is taken phenomenologically into account by the convolution with the line-width function (2.8) in (2.6). In the semiclassical limit "$\hbar \to 0$," $\Gamma \to 0$ and we recover the semiclassical Fermi–Dirac distribution.

The WPBGK system (2.1) to (2.4) should be solved for a Wigner function which is $2\pi/l$-periodic in k and satisfies appropriate initial and boundary conditions. It is convenient to derive the charge continuity equation and a nonlocal Ampère's law for the current density. The Wigner function f is periodic in k; its Fourier expansion is

$$f(x,k,t) = \sum_{j=-\infty}^{\infty} f_j(x,t)\, e^{ijkl}. \tag{2.10}$$

Defining $F = \partial W/\partial x$ (*minus* the electric field) and the average

$$\langle F\rangle_j(x,t) = \frac{1}{jl}\int_{-jl/2}^{jl/2} F(x+s,t)\, ds, \tag{2.11}$$

it is possible to obtain the following equivalent form of the Wigner equation [2]:

$$\frac{\partial f}{\partial t} + \sum_{j=-\infty}^{\infty} \frac{ijl}{\hbar}\, e^{ijkl} \left(\mathcal{E}_j \frac{\partial}{\partial x}\langle f\rangle_j + e\,\langle F\rangle_j\, f_j \right) = Q[f]. \tag{2.12}$$

Here the nonzero Fourier coefficients of the dispersion relation are simply $\mathcal{E}_0 = \Delta/2$ and $\mathcal{E}_{\pm 1} = -\Delta/4$ for the tight-binding dispersion relation $\mathcal{E}(k) = \Delta\,(1 - \cos kl)/2$ (Δ is the miniband width), which yields a miniband group velocity $v(k) = \frac{\Delta l}{2\hbar}\sin kl$. Integrating this equation over k yields the charge continuity equation

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x} \sum_{j=1}^{\infty} \frac{2jl}{\hbar}\, \langle \mathrm{Im}(\mathcal{E}_{-j} f_j)\rangle_j = 0. \tag{2.13}$$

Here we can eliminate the electron density by using the Poisson equation and then integrate over x, thereby obtaining the nonlocal Ampère's law for the total current density $J(t)$:

$$\varepsilon\,\frac{\partial F}{\partial t} + \frac{2e}{\hbar} \sum_{j=1}^{\infty} j\langle \mathrm{Im}(\mathcal{E}_{-j} f_j)\rangle_j = J(t). \tag{2.14}$$

To derive the quantum drift-diffusion equation, we shall assume that the electric field contribution in (2.12) is comparable to the collision terms and that they dominate the other terms (*the hyperbolic limit*) [5]. Let v_M and F_M be the electron velocity and field positive values at which the (zeroth order) drift velocity reaches its maximum. In this limit, the time t_0 it takes an electron with speed v_M to traverse a distance $x_0 = \varepsilon F_M l/(eN_D)$, over which the field variation is of order F_M, is much longer than the mean free time between collisions, $\nu_e^{-1} \sim \hbar/(eF_M l) = t_1$. We therefore define the *small parameter* $\lambda = t_1/t_0 = \hbar v_M N_D/(\varepsilon F_M^2 l^2)$ and formally multiply the first two terms on the left side of (2.1) or (2.12) by λ [5, 2]. The result is

$$\lambda\left(\frac{\partial f}{\partial t} + \sum_{j=-\infty}^{\infty} \frac{ijl}{\hbar}\, e^{ijkl}\mathcal{E}_j \frac{\partial}{\partial x}\langle f\rangle_j \right) = Q[f] - \sum_{j=-\infty}^{\infty} \frac{iejl}{\hbar}\, e^{ijkl}\langle F\rangle_j\, f_j. \tag{2.15}$$

The solution of (2.15) for $\lambda = 0$ is calculated in terms of its Fourier coefficients as

$$f^{(0)}(k;F) = \sum_{j=-\infty}^{\infty} \frac{(1 - ij\mathcal{F}_j/\tau_e)\, f_j^{FD}}{1 + j^2\mathcal{F}_j^2}\, e^{ijkl}, \tag{2.16}$$

where $\mathcal{F}_j = \langle F\rangle_j/F_M$, $F_M = \frac{\hbar}{el}\sqrt{\nu_e(\nu_e + \nu_i)}$ and $\tau_e = \sqrt{(\nu_e + \nu_i)/\nu_e}$.

The Chapman–Enskog ansatz for the Wigner function is [2]

$$f(x,k,t;\lambda) = f^{(0)}(k;F) + \sum_{m=1}^{\infty} f^{(m)}(k;F)\,\lambda^m, \tag{2.17}$$

$$\varepsilon \frac{\partial F}{\partial t} + \sum_{m=0}^{\infty} J^{(m)}(F)\,\lambda^m = J(t). \tag{2.18}$$

The coefficients $f^{(m)}(k;F)$ depend on the "slow variables" x and t only through their dependence on the electric field and the electron density. The electric field obeys a reduced evolution equation (2.18) in which the functionals $J^{(m)}(F)$ are chosen so that the $f^{(m)}(k;F)$ are bounded and $2\pi/l$-periodic in k. After we keep the desired number of terms and set $\lambda = 1$, (2.18) is the quantum drift-diffusion equation provided by our perturbation procedure.

Differentiating Ampère's law (2.18) with respect to x, we obtain the charge continuity equation. Moreover the compatibility condition

$$\int_{-\pi/l}^{\pi/l} f^{(m)}(k;n)\,dk = \frac{2\pi}{l}\, f_0^{(m)} = 0, \quad m \ge 1, \tag{2.19}$$

is obtained by inserting the expansion (2.17) into (2.3). Inserting (2.17) and (2.18) in (2.15), we find the hierarchy

$$\mathcal{L}f^{(1)} = -\left.\frac{\partial f^{(0)}}{\partial t}\right|_0 + \sum_{j=-\infty}^{\infty} \frac{ijl\mathcal{E}_j e^{ijkl}}{\hbar}\frac{\partial}{\partial x}\langle f^{(0)}\rangle_j, \tag{2.20}$$

$$\mathcal{L}f^{(2)} = -\left.\frac{\partial f^{(1)}}{\partial t}\right|_0 + \sum_{j=-\infty}^{\infty} \frac{ijl\mathcal{E}_j e^{ijkl}}{\hbar}\frac{\partial}{\partial x}\langle f^{(1)}\rangle_j \; - \left.\frac{\partial}{\partial t} f^{(0)}\right|_1, \tag{2.21}$$

and so on. Here

$$\mathcal{L}u(k) \equiv \frac{ie}{\hbar}\sum_{-\infty}^{\infty} jl\langle F\rangle_j u_j e^{ijkl} + \left(\nu_e + \frac{\nu_i}{2}\right) u(k) - \frac{\nu_i}{2}\, u(-k), \tag{2.22}$$

and the subscripts 0 and 1 on the right-hand side of these equations mean that $\varepsilon\,\partial F/\partial t$ is replaced by $J - J^{(0)}(F)$ and by $-J^{(1)}(F)$, respectively.

The condition (2.19) implies that

$$\int_{-\pi/l}^{\pi/l} \mathcal{L}f^{(m)} dk = 0 \tag{2.23}$$

for $m \ge 1$. Using this, the solvability conditions for the linear hierarchy of equations yield

$$J^{(m)} = \frac{2e}{\hbar}\sum_{j=1}^{\infty} j\langle \mathrm{Im}(\mathcal{E}_{-j} f_j^{(m)})\rangle_j, \tag{2.24}$$

which can also be obtained by insertion of (2.17) into (2.14).

Particularized to the case of the tight-binding dispersion relation and $\Gamma = 0$ in the Fermi–Dirac distribution (2.4), the leading order of the Ampère law (2.18) is

$$\varepsilon \frac{\partial F}{\partial t} + \frac{ev_M}{l} \langle n\mathcal{M}V(\mathcal{F})\rangle_1 = J(t), \tag{2.25}$$

$$V(\mathcal{F}) = \frac{2\mathcal{F}}{1+\mathcal{F}^2}, \quad v_M = \frac{\Delta l\, \mathcal{I}_1(M)}{4\hbar\tau_e \mathcal{I}_0(M)}, \quad \mathcal{M}\left(\frac{n}{N_D}\right) = \frac{\mathcal{I}_1(\tilde{\mu})\,\mathcal{I}_0(M)}{\mathcal{I}_1(M)\,\mathcal{I}_0(\tilde{\mu})}, \tag{2.26}$$

$$\mathcal{I}_m(s) = \int_{-\pi}^{\pi} \cos(mk)\, \ln\left(1 + e^{s-\delta+\delta\cos k}\right) dk, \tag{2.27}$$

provided $\mathcal{F} \equiv \mathcal{F}_1$, $\delta = \Delta/(2k_BT)$, and $\tilde{\mu} \equiv \mu/(k_BT)$. Here M (calculated graphically in Figure 1 of [5]) is the value of the dimensionless chemical potential $\tilde{\mu}$ at which (2.3) holds with $n = N_D$. The drift velocity $v_M V(\mathcal{F})$ has the Esaki–Tsu form with a peak velocity that becomes $v_M \approx \Delta l I_1(\delta)/[4\hbar\tau_e I_0(\delta)]$ in the Boltzmann limit [11] ($I_n(\delta)$ is the modified Bessel function of the nth order).

To find the first-order correction in (2.18), we first solve (2.20) and find $J^{(m)}$ for $m = 1$. The calculation yields the first correction to (2.25) (here $'$ means differentiation with respect to n) [2]:

$$\varepsilon \frac{\partial F}{\partial t} + \frac{ev_M}{l} \mathcal{N}\left(F, \frac{\partial F}{\partial x}\right) = \varepsilon \left\langle D\left(F, \frac{\partial F}{\partial x}, \frac{\partial^2 F}{\partial x^2}\right)\right\rangle_1 + \langle A\rangle_1\, J(t), \tag{2.28}$$

$$A = 1 + \frac{2ev_M}{\varepsilon F_M l(\nu_e+\nu_i)}\, \frac{1-(1+2\tau_e^2)\,\mathcal{F}^2}{(1+\mathcal{F}^2)^3}\, n\mathcal{M}, \tag{2.29}$$

$$\mathcal{N} = \langle nV\mathcal{M}\rangle_1 + \langle (A-1)\langle\langle nV\mathcal{M}\rangle_1\rangle_1\rangle_1 - \frac{\Delta l\tau_e}{F_M\hbar(\nu_e+\nu_i)} \left\langle \frac{B}{1+\mathcal{F}^2}\right\rangle_1, \tag{2.30}$$

$$D = \frac{\Delta^2 l^2}{8\hbar^2(\nu_e+\nu_i)(1+\mathcal{F}^2)} \left(\frac{\partial^2\langle F\rangle_1}{\partial x^2} - \frac{4\hbar v_M\tau_e C}{\Delta l}\right), \tag{2.31}$$

$$\begin{aligned} B = {} & \left\langle \frac{4\mathcal{F}_2 n\mathcal{M}_2}{(1+4\mathcal{F}_2^2)^2} \frac{\partial\langle F\rangle_2}{\partial x}\right\rangle_1 + \mathcal{F}\left\langle \frac{n\mathcal{M}_2(1-4\mathcal{F}_2^2)}{(1+4\mathcal{F}_2^2)^2}\frac{\partial\langle F\rangle_2}{\partial x}\right\rangle_1 \\ & - \frac{4\hbar v_M(1+\tau_e^2)\mathcal{F}(n\mathcal{M})'}{\Delta l\tau_e(1+\mathcal{F}^2)} \left\langle n\mathcal{M}\frac{1-\mathcal{F}^2}{(1+\mathcal{F}^2)^2}\,\frac{\partial\langle F\rangle_1}{\partial x}\right\rangle_1, \end{aligned} \tag{2.32}$$

$$\begin{aligned} C = {} & \left\langle \frac{(n\mathcal{M}_2)'}{1+4\mathcal{F}_2^2}\, \frac{\partial^2 F}{\partial x^2}\right\rangle_1 - 2\mathcal{F}\left\langle \frac{(n\mathcal{M}_2)'\mathcal{F}_2}{1+4\mathcal{F}_2^2}\, \frac{\partial^2 F}{\partial x^2}\right\rangle_1 \\ & + \frac{8\hbar v_M(1+\tau_e^2)(n\mathcal{M})'\,\mathcal{F}}{\Delta l\tau_e\,(1+\mathcal{F}^2)} \left\langle \frac{(n\mathcal{M})'\mathcal{F}}{1+\mathcal{F}^2}\, \frac{\partial^2 F}{\partial x^2}\right\rangle_1. \end{aligned} \tag{2.33}$$

Here $\mathcal{M}_2(n/N_D) \equiv \mathcal{I}_2(\tilde{\mu})\,\mathcal{I}_0(M)/[\mathcal{I}_1(M)\,\mathcal{I}_0(\tilde{\mu})]$. If the electric field and the electron density do not change appreciably over two SL periods, $\langle F\rangle_j \approx F$, the spatial averages can be ignored, and the *nonlocal* quantum drift-diffusion equation (2.28) becomes the *local* generalized drift-diffusion equation obtained from the semiclassical theory [5]. The boundary conditions for the quantum drift-diffusion equation (2.28) (which contains triple spatial averages) need to be specified on the intervals $[-2l, 0]$ and $[Nl, Nl + 2l]$, not just at the points $x = 0$ and $x = Nl$, as in the case of the parabolic generalized drift-diffusion equation. Similarly, the initial condition has to be defined on the extended interval $[-2l, Nl + 2l]$. For realistic values of the parameters representing a strongly coupled SL under dc voltage bias, the numerical solution of the quantum drift-diffusion equation yields a stable self-sustained oscillation of the current [2] in quantitative agreement with experiments [20]. Details of the numerical procedure can be found in [9].

3. Wigner description of a two-miniband superlattice. We shall consider a 2×2 Hamiltonian $\mathbf{H}(x, -i\partial/\partial x)$, in which [13]

$$\begin{aligned} \mathbf{H}(x,k) &= [h_0(k) - eW(x)]\boldsymbol{\sigma}_0 + \vec{h}(k)\cdot\vec{\boldsymbol{\sigma}}] \\ &\equiv \begin{pmatrix} (\alpha+\gamma)(1-\cos kl) - eW(x) + g & -i\beta\sin kl \\ i\beta\sin kl & (\alpha-\gamma)(1-\cos kl) - eW(x) - g \end{pmatrix}. \end{aligned} \tag{3.1}$$

Here

$$\begin{aligned} h_0(k) &= \alpha\,(1-\cos kl), & h_1(k) &= 0, \\ h_2(k) &= \beta\sin kl, & h_3(k) &= \gamma\,(1-\cos kl) + g, \end{aligned} \tag{3.2}$$

and

$$\boldsymbol{\sigma}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{3.3}$$

are the Pauli matrices.

The Hamiltonian (3.1) corresponds to the simplest 2×2 Kane model in which the quadratic and linear terms $(kl)^2/2$ and kl are replaced by $(1-\cos kl)$ and $\sin kl$, respectively. For an SL with two minibands, $2g$ is the miniband gap and $\alpha = (\Delta_1 + \Delta_2)/4$ and $\gamma = (\Delta_1 - \Delta_2)/4$, provided Δ_1 and Δ_2 are the miniband widths. In the case of an LSL, $g = \gamma = 0$, and $h_2\boldsymbol{\sigma}_2$ corresponds to the precession term in the Rashba spin-orbit interaction [14]. The other term, the intersubband coupling, depends on the momentum in the y direction, and we have not included it here. Small modifications of (3.1) represent a single miniband SL with dilute magnetic impurities in the presence of a magnetic field B: $g = \gamma = h_2 = 0$, and $h_1 = \beta(B)$ [19]. As in the case of a single miniband SL, $W(x)$ is the electric potential.

The energy minibands $\mathcal{E}^{\pm}(k)$ are the eigenvalues of the free Hamiltonian $\mathbf{H}_0(k) = h_0(k)\boldsymbol{\sigma}_0 + \vec{h}(k)\cdot\vec{\boldsymbol{\sigma}}$ and are given by

$$\mathcal{E}^{\pm}(k) = h_0(k) \pm |\vec{h}(k)|. \tag{3.4}$$

The corresponding spectral projections are

$$\mathbf{P}^{\pm}(k) = \frac{\boldsymbol{\sigma}_0 \pm \vec{\nu}(k)\cdot\vec{\boldsymbol{\sigma}}}{2}, \quad \text{where} \quad \vec{\nu}(k) = \vec{h}(k)/|\vec{h}(k)|, \tag{3.5}$$

so that we can write

$$\mathbf{H}_0(k) = \mathcal{E}^{+}(k)\mathbf{P}^{+}(k) + \mathcal{E}^{-}(k)\mathbf{P}^{-}(k). \tag{3.6}$$

We shall now write the WPBGK equations for the Wigner matrix written in terms of the Pauli matrices:

$$\mathbf{f}(x,k,t) = \sum_{i=0}^{3} f^i(x,k,t)\boldsymbol{\sigma}_i = f^0(x,k,t)\boldsymbol{\sigma}_0 + \vec{f}(x,k,t)\cdot\vec{\boldsymbol{\sigma}}. \tag{3.7}$$

The Wigner components are real and can be related to the coefficients of the Hermitian Wigner matrix by

$$\begin{aligned} f_{11} &= f^0 + f^3, & f_{12} &= f^1 - if^2, \\ f_{21} &= f^1 + if^2, & f_{22} &= f^0 - f^3. \end{aligned} \tag{3.8}$$

Hereinafter we shall use the equivalent notation

$$f = \begin{pmatrix} f^0 \\ \vec{f} \end{pmatrix} = \begin{pmatrix} f^0 \\ f^1 \\ f^2 \\ f^3 \end{pmatrix}. \tag{3.9}$$

The populations of the minibands with energies $\mathcal{E}^{\pm}$ are given by the moments

$$n^{\pm}(x,t) = \frac{l}{2\pi} \int_{-\pi/l}^{\pi/l} \left[f^0(x,k,t) \pm \vec{\nu}(k) \cdot \vec{f}(x,k,t) \right] dk, \tag{3.10}$$

and the total electron density is $n^+ + n^-$. After some algebra, we can obtain the following WPBGK equations for the Wigner components:

$$\frac{\partial f^0}{\partial t} + \frac{\alpha}{\hbar} \sin kl \, \Delta^- f^0 + \vec{b} \cdot \Delta^- \vec{f} - \Theta f^0 = Q^0[f], \tag{3.11}$$

$$\frac{\partial \vec{f}}{\partial t} + \frac{\alpha}{\hbar} \sin kl \, \Delta^- \vec{f} + \vec{b} \, \Delta^- f^0 + \vec{\omega} \times \vec{f} - \Theta \vec{f} = \vec{Q}[f], \tag{3.12}$$

$$\varepsilon \, \frac{\partial^2 W}{\partial x^2} = \frac{e}{l} \, (n^+ + n^- - N_D), \tag{3.13}$$

whose right-hand sides contain collision terms to be described later. Here

$$(\Delta^{\pm} u)(x,k) = u(x + l/2, k) \pm u(x - l/2, k), \tag{3.14}$$

$$\vec{\omega} = \vec{\omega}_0 + \vec{\omega}_1, \tag{3.15}$$

$$\vec{\omega}_0 = \frac{2g}{\hbar} \, (0, 0, 1), \tag{3.16}$$

$$\vec{\omega}_1 = \frac{1}{\hbar} \, (0, \beta \sin kl \, \Delta^+, 2\gamma - \gamma \cos kl \, \Delta^+), \tag{3.17}$$

$$\vec{b} = \frac{1}{\hbar} \, (0, \beta \cos kl, \gamma \sin kl), \tag{3.18}$$

$$\Theta f^i(x,k,t) = \sum_{j=-\infty}^{\infty} \frac{ejl}{i\hbar} \langle F(x,t) \rangle_j e^{ijkl} f^i_j(x,t). \tag{3.19}$$

Our collision model contains two terms: a BGK term which tries to send the miniband Wigner function to its local equilibrium and a scattering term from the miniband with higher energy to the lowest miniband:

$$Q^0[f] = -\frac{f^0 - \Omega^0}{\tau}, \tag{3.20}$$

$$\vec{Q}[f] = -\frac{\vec{f} - \vec{\Omega}}{\tau} - \frac{\vec{\nu} f^0 + \vec{f}}{\tau_{\mathrm{sc}}}, \tag{3.21}$$

$$\Omega^0 = \frac{\phi^+ + \phi^-}{2}, \quad \vec{\Omega} = \frac{\phi^+ - \phi^-}{2} \, \vec{\nu}, \tag{3.22}$$

$$\phi^{\pm}(k; n^{\pm}) = \frac{m^* k_B T}{\pi \hbar^2} \int_{-\infty}^{\infty} \frac{\sqrt{2}\, \Gamma^3/\pi}{\Gamma^4 + [E - \mathcal{E}^{\pm}(k)]^4} \, \ln\left(1 + e^{\frac{\mu^{\pm} - E}{k_B T}}\right) dE, \tag{3.23}$$

$$n^{\pm} = \frac{l}{2\pi} \int_{-\pi/l}^{\pi/l} \phi^{\pm}(k; n^{\pm}) \, dk. \tag{3.24}$$

The chemical potentials of the minibands, μ^+ and μ^-, are calculated in terms of n^+ and n^-, respectively, by inserting (3.23) in (3.24) and solving the resulting equations. Our collision model should enforce charge continuity. To check this, we first calculate the time derivative of $n^\pm$ using (3.10) to (3.12):

$$\begin{aligned}(3.25)\quad &\frac{\partial n^\pm}{\partial t} + \frac{\alpha l \Delta^-}{2\pi\hbar}\int_{-\pi/l}^{\pi/l} \sin kl\,(f^0 \pm \vec{\nu}\cdot\vec{f})\,dk + \frac{l\Delta^-}{2\pi}\int_{-\pi/l}^{\pi/l}(\vec{b}\cdot\vec{f} \pm \vec{\nu}\cdot\vec{b}f^0)\,dk\\ &\pm\frac{l\Delta^-}{2\pi}\int_{-\pi/l}^{\pi/l}\vec{\nu}\cdot\vec{\omega}\times\vec{f}\,dk \mp \frac{l\Delta^-}{2\pi}\int_{-\pi/l}^{\pi/l}\vec{\nu}\cdot\Theta\vec{f}\,dk\\ &= \frac{l\Delta^-}{2\pi}\int_{-\pi/l}^{\pi/l}(Q^0[f] \pm \vec{\nu}\cdot\vec{Q}[f])\,dk = \mp\frac{n^+}{\tau_{\rm sc}},\end{aligned}$$

where we have employed $\int \Theta f^0 dk = 0$. Then we obtain

$$(3.26)\qquad \frac{\partial}{\partial t}(n^+ + n^-) + \Delta^-\left[\frac{l}{\pi}\int_{-\pi/l}^{\pi/l}\left(\frac{\alpha}{\hbar}\sin kl\, f^0 + \vec{b}\cdot\vec{f}\right)dk\right] = 0.$$

Noting that $\Delta^- u(x) = l\,\partial\langle u(x)\rangle_1/\partial x$, we see that this equation corresponds to charge continuity. Differentiating in time the Poisson equation (3.13), using (3.26) in the result, and integrating with respect to x, we get the following nonlocal Ampère law for the balance of current:

$$(3.27)\qquad \varepsilon\frac{\partial F}{\partial t} + \left\langle \frac{el}{\pi}\int_{-\pi/l}^{\pi/l}\left(\frac{\alpha}{\hbar}\sin kl\, f^0 + \vec{b}\cdot\vec{f}\right)dk\right\rangle_1 = J(t).$$

Here the space-independent function $J(t)$ is the total current density. Since the Wigner components are real, we can rewrite (3.27) in the following equivalent form:

$$(3.28)\qquad \varepsilon\frac{\partial F}{\partial t} - \frac{2e}{\hbar}\left\langle \alpha\,\mathrm{Im} f_1^0 - \beta\,\mathrm{Re} f_1^2 + \gamma\,\mathrm{Im} f_1^3\right\rangle_1 = J(t).$$

4. Derivation of balance equations by the Chapman–Enskog method.

In this section, we shall derive the reduced balance equations for our two-miniband SL using the Chapman–Enskog method. First of all, we should decide the order of magnitude of the terms in the WPBGK equations (3.11) and (3.12) in the hyperbolic limit. Recall that in this limit, the collision frequency $1/\tau$ and the Bloch frequency $eF_M l/\hbar$ are of the same order, about 10 THz for the SL of section 2. Typically, $2g/\hbar$ is of the same order, so that the term containing $\vec{\omega}_0$ should also balance the BGK collision term. What about the other terms?

The scattering time $\tau_{\rm sc}$ is much longer than the collision time τ, and we shall consider $\tau/\tau_{\rm sc} = O(\lambda) \ll 1$. Moreover, the gap energy is typically much larger than the miniband widths or the spin-orbit coefficient, and a rich dominant balance is obtained by assuming that β/g and γ/g are of order λ. Then we can expand the unit vector $\vec{\nu}$ as follows:

$$\begin{aligned}(4.1)\qquad \vec{\nu} = (0,0,1) + \frac{\lambda\beta}{g}\sin kl\,(0,1,0) - \lambda^2\Bigg[&\frac{\beta\gamma}{g^2}\sin kl(1-\cos kl)\,(0,1,0)\\ &+\frac{\beta^2\sin^2 kl}{2g^2}\,(0,0,1)\Bigg] + O(\lambda^3).\end{aligned}$$

In this expansion, we have inserted the bookkeeping parameter λ, which is set equal to 1 at the end of our calculations (cf. section 2). From (3.11) and (3.12), we can write the scaled WPBGK equations as follows:

$$\mathbb{L}f - \Omega = -\lambda \left(\tau \frac{\partial f}{\partial t} + \Lambda f \right). \tag{4.2}$$

Here the operators $\mathbb{L}$ and Λ are defined by

$$\mathbb{L}f = f - \tau\,\Theta f + \delta_1 \begin{pmatrix} 0 \\ -f^2 \\ f^1 \\ 0 \end{pmatrix}, \tag{4.3}$$

$$\Lambda f = \delta_2 \begin{pmatrix} 0 \\ \vec{f} + \vec{\nu} f^0 \end{pmatrix} + \frac{\alpha\tau}{\hbar} \sin kl\, \Delta^- f + \Delta^- \begin{pmatrix} \tau \vec{b}\cdot\vec{f} \\ \tau\,\vec{b}\, f^0 \end{pmatrix} + \begin{pmatrix} 0 \\ \tau\,\vec{\omega}_1 \times \vec{f} \end{pmatrix}, \tag{4.4}$$

where

$$\delta_1 = \frac{2g\tau}{\hbar}, \quad \delta_2 = \frac{\tau}{\tau_{\mathrm{sc}}}. \tag{4.5}$$

The expansion of $\vec{\nu}$ in powers of λ gives rise to a similar expansion of Ω and Λ.

To derive the reduced balance equations, we use the following Chapman–Enskog ansatz:

$$f(x,k,t;\lambda) = f^{(0)}(k; n^+, n^-, F) + \sum_{m=1}^{\infty} f^{(m)}(k; n^+, n^-, F)\,\lambda^m, \tag{4.6}$$

$$\varepsilon \frac{\partial F}{\partial t} + \sum_{m=0}^{\infty} J_m(n^+, n^-, F)\,\lambda^m = J(t), \tag{4.7}$$

$$\frac{\partial n^{\pm}}{\partial t} = \sum_{m=0}^{\infty} A_m^{\pm}(n^+, n^-, F)\,\lambda^m. \tag{4.8}$$

The functions $A_m^{\pm}$ and J_m are related through the Poisson equation (3.13), so that

$$A_m^+ + A_m^- = -\frac{l}{e}\frac{\partial J_m}{\partial x}. \tag{4.9}$$

Inserting (4.6) to (4.8) into (4.2), we get

$$\mathbb{L}f^{(0)} = \Omega_0, \tag{4.10}$$

$$\mathbb{L}f^{(1)} = \Omega_1 - \tau \left.\frac{\partial f^{(0)}}{\partial t}\right|_0 - \Lambda_0 f^{(0)}, \tag{4.11}$$

$$\mathbb{L}f^{(2)} = \Omega_2 - \tau \left.\frac{\partial f^{(1)}}{\partial t}\right|_0 - \Lambda_0 f^{(1)} - \tau \left.\frac{\partial f^{(0)}}{\partial t}\right|_1 - \Lambda_1 f^{(0)}, \tag{4.12}$$

and so on. The subscripts 0 and 1 on the right-hand side of these equations mean that we replace $\varepsilon\,\partial F/\partial t|_m = J\delta_{0m} - J_m$, $\partial n^{\pm}/\partial t|_m = A_m^{\pm}$. Moreover, inserting (4.1) and (4.6) into (3.10) yields the following compatibility conditions:

$$f_0^{(1)\,0} = 0, \quad f_0^{(1)\,3} = \frac{\beta}{g}\,\mathrm{Im} f_1^{(0)\,2}, \tag{4.13}$$

$$\begin{aligned} &f_0^{(2)\,0} = 0, \\ &f_0^{(2)\,3} = \frac{\beta}{g}\,\mathrm{Im} f_1^{(1)\,2} + \frac{\beta^2}{4g^2}\,(f_0^{(0)\,3} - \mathrm{Re} f_2^{(0)\,3}) - \frac{\beta\gamma}{g^2}\,\mathrm{Im}\left(f_1^{(0)\,2} - \frac{f_2^{(0)\,2}}{2}\right), \end{aligned} \tag{4.14}$$

etc.

To solve (4.10) for $f^{(0)} \equiv \varphi$, we first note that

$$-\tau\,\Theta\varphi = \sum_{j=-\infty}^{\infty} i\vartheta_j \varphi_j e^{ijkl}, \tag{4.15}$$

$$\vartheta_j \equiv \frac{\tau e j l}{\hbar}\,\langle F\rangle_j. \tag{4.16}$$

Then (4.10), (3.22), and (4.1) yield

$$\varphi_j^0 = \frac{\phi_j^+ + \phi_j^-}{2}\,\frac{1 - i\vartheta_j}{1+\vartheta_j^2}, \quad \varphi_j^1 = \varphi_j^2 = 0, \quad \varphi_j^3 = \frac{\phi_j^+ - \phi_j^-}{2}\,\frac{1 - i\vartheta_j}{1+\vartheta_j^2}, \tag{4.17}$$

where we have used the fact that the Fourier coefficients

$$\phi_j^{\pm} = \frac{l}{\pi}\int_0^{\pi/l} \cos(jkl)\,\phi^{\pm}\,dk \tag{4.18}$$

are real because $\phi^{\pm}$ are even functions of k. Similarly, the solution of (4.11) is $f^{(1)} \equiv \psi$ with

$$\begin{aligned} \psi_j^m &= r_j^m\,\frac{1 - i\vartheta_j}{1+\vartheta_j^2} \quad (m = 0, 3), \\ \psi_j^1 &= \frac{(1+i\vartheta_j)\,r_j^1 + \delta_1\,r_j^2}{(1+i\vartheta_j)^2 + \delta_1^2}, \\ \psi_j^2 &= \frac{(1+i\vartheta_j)\,r_j^2 - \delta_1\,r_j^1}{(1+i\vartheta_j)^2 + \delta_1^2}. \end{aligned} \tag{4.19}$$

Here r is the right-hand side of (4.11). The balance equations can be found in two ways. We can calculate $A_m^{\pm}$ for $m = 0, 1$ by using the compatibility conditions (4.13) and (4.14) in (4.11) and (4.12), respectively. More simply, we can insert the solutions (4.17) and (4.19) in the balance equations (3.25) and in the Ampère law (3.27). The result is

$$\frac{\partial n^{\pm}}{\partial t} + \Delta^- D_{\pm}(n^+, n^-, F) = \pm R(n^+, n^-, F), \tag{4.20}$$

$$\begin{aligned} &\varepsilon\,\frac{\partial F}{\partial t} + \frac{e}{\hbar}\left\langle [\alpha\,(\phi_1^+ + \phi_1^-) + \gamma\,(\phi_1^+ - \phi_1^-)]\,\frac{\vartheta_1}{1+\vartheta_1^2}\right\rangle_1 \\ &\quad + \frac{2e}{\hbar}\,[\beta\mathrm{Re}\langle\psi_1^2\rangle_1 - \alpha\,\mathrm{Im}\langle\psi_1^0\rangle_1 - \gamma\,\mathrm{Im}\langle\psi_1^3\rangle_1] = J, \end{aligned} \tag{4.21}$$

$$D_{\pm} = \frac{\alpha \pm \gamma}{\hbar}\left[\frac{\phi_1^{\pm}\vartheta_1}{1+\vartheta_1^2} - \mathrm{Im}(\psi_1^0 \pm \psi_1^3)\right] + \frac{\beta}{\hbar}\,\mathrm{Re}\psi_1^2 \pm \frac{\beta^2\vartheta_2}{4g\hbar}\,\frac{\phi_2^+ + \phi_2^-}{1+\vartheta_2^2}, \tag{4.22}$$

$$R = -\frac{\delta_2 n^+}{\tau} - \frac{\beta^2\vartheta_2^2(\phi_2^+ - \phi_2^-)}{8g^2\tau(1+\vartheta_2^2)} + \frac{\beta}{g\tau}\,\vartheta_1\mathrm{Re}\psi_1^2 + \frac{\beta}{\hbar}\,(2 - \Delta^+)\mathrm{Im}\psi_1^1. \tag{4.23}$$

The appendix justifies this second and more direct method by showing that equivalent expressions are obtained from the compatibility conditions. Note that (4.21) can be obtained from (4.20) and the Poisson equation.

5. Spintronics: Quantum drift-diffusion equations for an LSL with Rashba spin-orbit interaction. In the simpler case of an LSL with the precession term of Rashba spin-orbit interaction (but no intersubband coupling), we can obtain explicit rate equations for $n^{\pm}$ by means of the Chapman–Enskog method. In the Hamiltonian (3.1), we have $\gamma = g = 0$, so that $h_3 = 0$ and $\vec{\nu} = (0, 1, 0)$. However, the Fermi–Dirac distribution is different from (2.6) for an LSL. We have to replace E_n instead of $\hbar^2 k_z^2/(2m^*)$, sum over n for all populated QW energy levels, and integrate over k_y only. Provided only E_1 is populated, we obtain the following expression instead of (3.23):

$$\phi^{\pm}(k; n^{\pm}) = \int_{-\infty}^{\infty} \frac{D_\Gamma\left(E - \mathcal{E}^{\pm}(k) - E_1\right)}{1 + \exp\left(\frac{E-\mu^{\pm}}{k_B T}\right)}\, dE, \tag{5.1}$$

where the broadened density of states is

$$D_\Gamma(E) = \frac{1}{2\pi L_z} \int_{-\infty}^{\infty} dk_y \delta_\Gamma \left(\frac{\hbar^2 k_y^2}{2m^*} - E\right) = \frac{\sqrt{2m^*}}{2\pi\hbar L_z} \int_0^{\infty} dE_y\, \frac{\delta_\Gamma(E_y - E)}{\sqrt{E_y}}. \tag{5.2}$$

Note that (5.2) becomes the 1D density of states $D(E) = \sqrt{2m^*}\theta(E)/(2\pi\hbar L_z\sqrt{E})$ as $\Gamma \to 0+$. We have not included a factor 2 in (5.2) because all the electrons in each of the minibands (with energies $\mathcal{E}^{\pm}(k)$) have the same spin. Inserting (2.8) in (5.2) and using the residue theorem to evaluate the integral, we obtain

$$\begin{aligned} D_\Gamma(E) = {} & \frac{\sqrt{m^*}}{4\pi\hbar L_z} \\ & \times \left[\frac{\sqrt{\sqrt{E^2 + \sqrt{2}\Gamma E + \Gamma^2} + E + \frac{\Gamma}{\sqrt{2}}} - \sqrt{\sqrt{E^2 + \sqrt{2}\Gamma E + \Gamma^2} - E - \frac{\Gamma}{\sqrt{2}}}}{\sqrt{E^2 + \sqrt{2}\Gamma E + \Gamma^2}} \right. \\ & \left. + \frac{\sqrt{\sqrt{E^2 - \sqrt{2}\Gamma E + \Gamma^2} + E - \frac{\Gamma}{\sqrt{2}}} + \sqrt{\sqrt{E^2 - \sqrt{2}\Gamma E + \Gamma^2} - E + \frac{\Gamma}{\sqrt{2}}}}{\sqrt{E^2 - \sqrt{2}\Gamma E + \Gamma^2}} \right]. \end{aligned} \tag{5.3}$$

As $E \to +\infty$, $D_\Gamma(E) \sim \sqrt{2m^*}/(2\pi\hbar L_z\sqrt{E})$, whereas $D_\Gamma(E) = O(|E|^{-5/2})$ as $E \to -\infty$. Then the convolution integral (5.1) is convergent.

In the present case, minibands correspond to electrons with spin up or down which have different energy. Scattering between minibands is the same as in (3.21), $-(\vec{\nu} f^0 + \vec{f})/\tau_{\text{sc}}$, which yields $\partial n^{\pm}/\partial t + \cdots = \mp n^{\pm}/\tau_{\text{sc}}$ in (3.25), only if the chemical potential of the miniband with lowest energy, μ^-, is less than the minimum energy of the other miniband, $\mathcal{E}^+_{\text{min}} = \min_k \mathcal{E}^+(k)$. Otherwise ($\mu^- > \mathcal{E}^+_{\text{min}}$), the scattering term should be $-2\vec{f}/\tau_{\text{sc}}$, which yields $\partial n^{\pm}/\partial t + \cdots = \mp(n^+ - n^-)/\tau_{\text{sc}}$ in (3.25), thereby trying to equalize n^+ and n^-; cf. [19].

Now we shall derive the balance equations in the hyperbolic limit using the Chapman–Enskog method as in section 4. In the scaled WPBGK equations (4.2), the operators $\mathbb{L}$ and Λ are

$$\mathbb{L}f = f - \tau\,\Theta f, \tag{5.4}$$

$$\Lambda f = \delta_2 \begin{pmatrix} 0 \\ 2\vec{f} + (\vec{\nu} f^0 - \vec{f})\,\theta(\mathcal{E}^+_{\min} - \mu^-) \end{pmatrix} + \frac{\alpha\tau}{\hbar}\sin kl\,\Delta^- f + \frac{\beta\tau}{\hbar}\cos kl\,\Delta^- \begin{pmatrix} f^2 \\ 0 \\ f^0 \\ 0 \end{pmatrix} + \frac{\beta\tau}{\hbar}\sin kl\,\Delta^+ \begin{pmatrix} 0 \\ f^3 \\ 0 \\ -f^1 \end{pmatrix}, \tag{5.5}$$

where δ_2 is given by (4.5), $\theta(x)$ is the Heaviside unit step function, and $\Omega^0 = (\phi^+ + \phi^-)/2$, $\vec{\Omega} = (0,1,0)\,(\phi^+ - \phi^-)/2$. The hierarchy of equations (4.10)–(4.12) is simply

$$\mathbb{L}f^{(0)} = \Omega, \tag{5.6}$$

$$\mathbb{L}f^{(1)} = -\,\tau\,\frac{\partial f^{(0)}}{\partial t}\bigg|_0 - \Lambda f^{(0)}, \tag{5.7}$$

$$\mathbb{L}f^{(2)} = -\,\tau\,\frac{\partial f^{(1)}}{\partial t}\bigg|_0 - \Lambda f^{(1)} - \tau\,\frac{\partial f^{(0)}}{\partial t}\bigg|_1, \tag{5.8}$$

and so on. The compatibility and solvability conditions are

$$f_0^{(m)\,0} = f_0^{(m)\,2} = 0 \quad \Longrightarrow (\mathbb{L}f^{(m)\,0})_0 = (\mathbb{L}f^{(m)\,2})_0 = 0, \quad m \geq 1. \tag{5.9}$$

The solution $f^{(0)} \equiv \varphi$ of (5.6) is

$$\varphi_j^0 = \frac{\phi_j^+ + \phi_j^-}{2}\,\frac{1 - i\vartheta_j}{1 + \vartheta_j^2}, \quad \varphi_j^1 = \varphi_j^3 = 0, \quad \varphi_j^2 = \frac{\phi_j^+ - \phi_j^-}{2}\,\frac{1 - i\vartheta_j}{1 + \vartheta_j^2}, \tag{5.10}$$

where we have used the fact that the Fourier coefficients $\phi_j^\pm$ are real because $\phi^\pm$ are even functions of k. Similarly, the solution of (5.7) is $f^{(1)} \equiv \psi$ with

$$\psi_j^m = r_j^m\,\frac{1 - i\vartheta_j}{1 + \vartheta_j^2} \quad (m = 0, 2), \quad \psi_j^1 = \psi_j^3 = 0. \tag{5.11}$$

Here r is the right-hand side of (5.7). The balance equations can be found in two ways. We can calculate $A_m^\pm$ for $m = 0, 1$ by using the solvability conditions (5.9) in (5.7) and (5.8), respectively. More simply, we can insert the solutions (5.10) and (5.11) in the balance equations (3.25) and in the Ampère law (3.27). In both cases, the result is

$$\frac{\partial n^\pm}{\partial t} + \Delta^- D_\pm(n^+, n^-, F) = \mp R(n^+, n^-, F), \tag{5.12}$$

$$\varepsilon\,\frac{\partial F}{\partial t} + e\,\langle D_+ + D_- \rangle_1 = J, \tag{5.13}$$

$$D_\pm = -\frac{\alpha}{\hbar}\,\Delta^- \mathrm{Im}(\varphi_1^0 \pm \varphi_1^2 + \psi_1^0 \pm \psi_1^2) \pm \frac{\beta}{\hbar}\,\Delta^- \mathrm{Re}(\varphi_1^0 \pm \varphi_1^2 + \psi_1^0 \pm \psi_1^2), \tag{5.14}$$

$$R = \frac{n^+ - n^-\,\theta(\mu^- - \mathcal{E}^+_{\min})}{\tau_{sc}}. \tag{5.15}$$

A straightforward calculation of (5.14) yields

$$
\begin{aligned}
D_{\pm} = {} & \frac{(\alpha\vartheta_1 \pm \beta)\phi_1^{\pm}}{\hbar\,(1+\vartheta_1^2)} \mp \frac{\tau\,(\phi_1^{+}-\phi_1^{-})\,[2\alpha\vartheta_1 \pm \beta(1-\vartheta_1^2)]}{2\hbar\tau_{\text{sc}}(1+\vartheta_1^2)^2} \\
& + \frac{[2\alpha\vartheta_1 \pm \beta(1-\vartheta_1^2)]\alpha\tau}{\hbar^2(1+\vartheta_1^2)^2}\,\frac{\partial\phi_1^{\pm}}{\partial n^{\pm}}\left[\Delta^{-}\left(\frac{\alpha\vartheta_1 \pm \beta}{\hbar\,(1+\vartheta_1^2)}\phi_1^{\pm}\right) \pm \frac{\hbar}{\alpha\tau_{sc}}(n^{+}-n^{-})\right] \\
& + \frac{\alpha\,(3\vartheta_1^2-1) \pm \beta\vartheta_1(3-\vartheta_1^2)}{\hbar(1+\vartheta_1^2)^3}\,\frac{l\tau^2\phi_1^{\pm}}{\hbar\varepsilon}\left(\frac{J}{e} - \left\langle\left\langle \frac{\alpha\,(\phi_1^{+}+\phi_1^{-})\vartheta_1}{\hbar(1+\vartheta_1^2)}\right\rangle_1\right\rangle_1\right. \\
& \left. - \left\langle\left\langle \frac{\beta\,(\phi_1^{+}-\phi_1^{-})}{\hbar(1+\vartheta_1^2)}\right\rangle_1\right\rangle_1\right) - \frac{(\alpha^2+\beta^2)\tau}{2\hbar^2(1+\vartheta_1^2)}\,\Delta^{-}n^{\pm} \\
& + \frac{\tau}{2\hbar^2(1+\vartheta_1^2)}\,\left[(\alpha^2-\beta^2 \mp 2\alpha\beta\vartheta_1)\,\Delta^{-}\left(\frac{\phi_2^{\pm}}{1+\vartheta_2^2}\right)\right. \\
& \left. + [(\beta^2-\alpha^2)\vartheta_1 \mp 2\alpha\beta]\,\Delta^{-}\left(\frac{\vartheta_2\phi_2^{\pm}}{1+\vartheta_2^2}\right)\right].
\end{aligned} \tag{5.16}
$$

We have numerically solved the system of equations (5.12)–(5.16), with the following boundary conditions in the interval $-2l \leq x \leq 0$:

$$\varepsilon\,\frac{\partial F}{\partial t} + \sigma\,F = J, \tag{5.17}$$

$$n^{+} = n^{-} = \frac{N_D}{2}, \tag{5.18}$$

whereas in the collector $Nl \leq x \leq Nl + 2l$,

$$\frac{\partial n^{\pm}}{\partial x} = \frac{\partial F}{\partial x} = 0 \tag{5.19}$$

hold. We have used the following values of the parameters: $\alpha = \Delta_1/2 = 8$ meV, $\beta = 2.63$ meV, $d_W = 3.1$ nm, $d_B = 1.96$ nm, $l = d_W + d_B = 5.06$ nm, $L_z = 3.1$ nm, $T = 5$ K, $\tau = 5.56 \times 10^{-14}$ s, $\tau_{\text{sc}} = 5.56 \times 10^{-13}$ s, $N_D = 4.048 \times 10^{10}$ cm^{-2}, $m^* = (0.067 d_W + 0.15 d_B) m_0/l$, $V = 3$ V, $N = 110$. We have used a large conductivity of the injecting contact $\sigma = 11.78\,\Omega^{-1}\text{m}^{-1}$. With these values, we select the following units to present our results graphically: $F_M = \hbar/(el\tau) = 23.417$ kV/cm, $x_0 = \varepsilon F_M l/(eN_D) = 19.4$ nm, $t_0 = \hbar/\alpha = 0.082$ ps, $J_0 = \alpha e N_D/(2\hbar) = 3.94 \times 10^4$ A/cm^2.

Figure 5.1(b)–(d) illustrates the resulting stable self-sustained current oscillations. They are due to the periodic formation of a pulse of the electric field at the cathode $x = 0$ and its motion through the LSL. Figure 5.1(b) depicts the pulse when it is far from the contacts, and the corresponding spin polarization is shown in Figure 5.1(d). It is interesting to consider the influence of the broadening Γ and the Fermi–Dirac statistics on the oscillations. At high temperatures, Boltzmann statistics and a semiclassical approximation should provide a good description. The semiclassical approximation is equivalent to dropping all spatial averages in our previous formulas. Since $x_0 \gg l$, the effect of dropping spatial averages should be rather small. Using Boltzmann statistics yields explicit formulas for $\mu^{\pm}$ in terms of $n^{\pm}$. In fact, we only have to replace $e^{(\mu^{\pm}-E)/(k_BT)}$ instead of the 3D Fermi distribution $[1 + e^{(E-\mu^{\pm})/(k_BT)}]^{-1}$ in (5.1). Using the relation (3.24) between $n^{\pm}$ and $\phi^{\pm}$, we obtain

$$\phi^{\pm} = n^{\pm}\,\frac{\pi\,\exp\left(\frac{\alpha\,\cos kl \mp \beta\,|\sin kl|}{k_BT}\right)}{\int_0^{\pi} dK\,\exp\left(\frac{\alpha\,\cos K \mp \beta\,\sin K}{k_BT}\right)}, \tag{5.20}$$

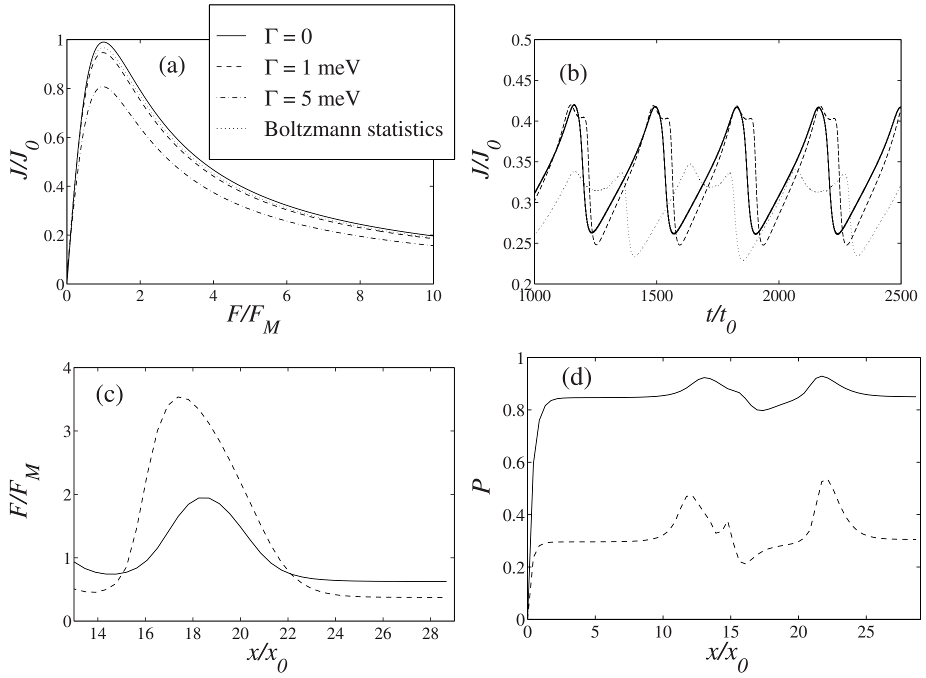


FIG. 5.1. (a) *Electron current vs. field in a spatially uniform stationary state for different values of the broadening* Γ *using the Fermi–Dirac distribution and for the Boltzmann distribution without broadening.* (b) *Total current density vs. time, and the* (c) *electric field and* (d) *spin polarization profiles during current self-oscillations for* $\Gamma = 0$ *(solid line) and* 1 *meV (dashed line). Parameter values are* $N = 110$, $N_D = 4.048 \times 10^{10}$ cm^{-2}, $d_B = 1.96$ *nm*, $L_z = d_W = 3.1$ *nm*, $l = 5.06$ *nm*, $\tau = 0.0556$ *ps*, $\tau_{\rm sc} = 0.556$ *ps*, $V = 3$ *V*, $\sigma = 11.78\,\Omega^{-1}m^{-1}$ $T = 5$ *K*,$m\alpha = 8$ *meV*, $\beta = 2.63$ *meV. With these values,* $\Delta_1 = 16$ *meV*, $x_0 = 19.4$ *nm*, $t_0 = 0.082$ *ps*, $J_0 = 3.94 \times 10^4$ A/cm^2.

and therefore,

$$\phi_j^{\pm} = n^{\pm} \frac{\int_0^{\pi} dK\, \cos(jK)\, \exp\left(\frac{\alpha \cos K \mp \beta \sin K}{k_B T}\right)}{\int_0^{\pi} dK\, \exp\left(\frac{\alpha \cos K \mp \beta \sin K}{k_B T}\right)} \tag{5.21}$$

for $j = 0, 1, \ldots$. Similar relations hold for the case of an SL with Boltzmann statistics in the tight-binding approximation.

The results are shown in Figure 5.1. Figure 5.1(a) depicts the relation between electron current and field for a spatially uniform stationary solution with $n^{\pm} = N_D/2$. We observe that all curves are similar. However, the curves for $\Gamma = 0$ and $\Gamma = 1$ meV are close, while the curve for $\Gamma = 5$ meV has dropped noticeably. The shapes of $J(t)$ for $\Gamma = 0$ and $\Gamma = 1$ meV in Figure 5.1(b) are close and quite different from that for $\Gamma = 5$ meV. If we look at the corresponding field profiles in Figure 5.1(c) and (d), for $\Gamma = 0$ and $\Gamma = 1$ meV the oscillations of the current are caused by the periodic nucleation of a pulse of the electric field at $x = 0$ and its motion towards the end of the LSL. The pulse far from the contacts shown in Figure 5.1(c) is larger in the case of $\Gamma = 0$ than for $\Gamma = 1$ meV. In the case of $\Gamma = 5$ meV (not shown), the pulse created at $x = 0$ becomes attenuated and disappears before arriving at $x = Nl$. This seems to indicate that the lowest voltage at which there exist stable self-sustained current oscillations is an increasing function of Γ: If we fix the voltage at 3 V and increase Γ, the critical voltage threshold to have stable oscillations approaches our fixed voltage of 3 V. Then the observed oscillations are smaller and the field profiles correspond to waves that vanish before reaching the end of the device, as it also occurs in models of the Gunn effect in bulk semiconductors [4].

6. Conclusions. We have presented a Wigner–Poisson–BGK system of equations with a collision broadened local Fermi–Dirac distribution for strongly coupled SLs having only one populated miniband. In the hyperbolic limit in which the collision and Bloch frequencies are of the same order and dominate all other frequencies, the Chapman–Enskog perturbation method yields a quantum drift-diffusion equation for the field. Numerical solutions of this equation exhibit self-sustained oscillations of the current due to recycling and motion of charge dipole domains [2].

For strongly coupled SLs having two populated minibands, we have introduced a periodic version of the Kane Hamiltonian and derived the corresponding Wigner–Poisson–BGK system of equations. The collision model comprises two terms, a BGK term trying to bring the Wigner matrix closer to a broadened Fermi–Dirac local equilibrium at each miniband, and a scattering term that brings down electrons from the upper to the lower miniband. By using the Chapman–Enskog method, we have derived quantum drift-diffusion equations for the miniband populations which contain generation-recombination terms. As it should be, the recombination terms vanish if there is no interminiband scattering and the off-diagonal terms in the Hamiltonian are zero. These terms may represent a Rashba spin-orbit interaction for an LSL. For an LSL under dc voltage bias in the growth direction, numerical solutions of the corresponding quantum drift-diffusion equations show self-sustained current oscillations due to periodic recycling and motion of electric field pulses. The periodic changes of the spin polarization and spin-polarized current indicate that this system acts as a spin oscillator.

Appendix. Balance equations from compatibility conditions We know that $\varphi^1 = \varphi^2 = 0$ from (4.11). Then the compatibility conditions (4.13) and (4.14) become

$$\psi_0^0 = 0, \quad \psi_0^3 = 0, \tag{A.1}$$

$$f_0^{(2)\,0} = 0, \quad f_0^{(2)\,3} = \frac{\beta}{g}\,\mathrm{Im}\psi_1^2 + \frac{\beta^2}{4g^2}\,(\varphi_0^3 - \mathrm{Re}\varphi_2^3). \tag{A.2}$$

Equations (A.1) imply that $(\mathbb{L}\psi)_0^m = 0$ for $m = 0, 3$ in (4.11). Since $\varphi_0^0 = (n^+ + n^-)/2$ and $\varphi_0^3 = (n^+ - n^-)/2$, these conditions yield

$$\begin{aligned}
&\frac{\tau}{2}\,\frac{\partial(n^+ + n^-)}{\partial t}\bigg|_0 - \frac{\alpha\tau}{\hbar}\,\Delta^-\mathrm{Im}\varphi_1^0 - \frac{\gamma\tau}{\hbar}\,\Delta^-\mathrm{Im}\varphi_1^3 = 0,\\
&\frac{\tau}{2}\,\frac{\partial(n^+ - n^-)}{\partial t}\bigg|_0 + \delta_2 n^+ - \frac{\alpha\tau}{\hbar}\,\Delta^-\mathrm{Im}\varphi_1^3 - \frac{\gamma\tau}{\hbar}\,\Delta^-\mathrm{Im}\varphi_1^0 = 0,
\end{aligned}$$

wherefrom we obtain

$$A_0^{\pm} = \mp\frac{n^+}{\tau_{\mathrm{sc}}} + \frac{\alpha \pm \gamma}{\hbar}\,\Delta^-\mathrm{Im}(\varphi_1^0 \pm \varphi_1^3). \tag{A.3}$$

Let us now calculate $A_1^\pm$. Equations (A.2) imply $(\mathbb{L}f^{(2)})_0^0 = 0$ and $(\mathbb{L}f^{(2)})_0^3 = f_0^{(2)\,3}$ given by (A.2) in (4.12). After a little algebra, we find

$$\begin{aligned} A_1^\pm = {}& \frac{\alpha \pm \gamma}{\hbar}\, \Delta^- \mathrm{Im}(\psi_1^0 \pm \psi_1^3) - \frac{\beta}{\hbar}\,(\Delta^- \mathrm{Re}\psi_1^2 \pm \Delta^+ \mathrm{Im}\psi_1^1) \\ & \mp \frac{\beta}{g\tau}\, \mathrm{Im}\psi_1^2 \pm \frac{\beta^2}{8g^2\tau}\,[2\mathrm{Re}\varphi_2^3 + \phi_2^+ - \phi_2^- - 2(n^+ - n^-)]. \end{aligned} \tag{A.4}$$

We will now transform (A.4) into an equivalent form by eliminating $\mathrm{Re}\varphi_2^3$ and $\mathrm{Im}\psi_1^2$ in favor of $\mathrm{Re}\varphi_2^3$ and $\mathrm{Im}\psi_1^2$, respectively. Equation (4.10) implies that $(1 + i\vartheta_2)\varphi_2^3 = (\phi_2^+ - \phi_2^-)/2$, and therefore

$$\mathrm{Re}\varphi_2^3 = \vartheta_2\, \mathrm{Im}\varphi_2^3 + \frac{\phi_2^+ - \phi_2^-}{2}. \tag{A.5}$$

Similarly, (4.11) implies that $(1 + i\vartheta_1)\,\psi_1^2 + \delta_1\,\psi_1^1 = r_1^2$, and therefore

$$\mathrm{Im}\psi_1^2 = -\vartheta_1\, \mathrm{Re}\psi_1^2 - \delta_1\, \mathrm{Im}\psi_1^1 + \mathrm{Im}r_1^2. \tag{A.6}$$

The right-hand side of (4.11) yields

$$r_1^2 = \frac{\beta}{2g}\left(\frac{1 - e^{-i2kl}}{2i}\,(\phi^+ - \phi^-)\right)_0 - \frac{\beta\tau}{\hbar}\,\Delta^-\left(\frac{1 + e^{-i2kl}}{2}\,\varphi^0\right)_0,$$

wherefrom

$$\mathrm{Im}r_1^2 = \frac{\beta}{4g}\,(\phi_2^+ - \phi_2^- - n^+ + n^-) - \frac{\beta\tau}{2\hbar}\,\Delta^- \mathrm{Im}\varphi_2^0. \tag{A.7}$$

Inserting (A.5), (A.6), and (A.7) into (A.4), we obtain the equivalent form

$$\begin{aligned} A_1^\pm = {}& \frac{\alpha \pm \gamma}{\hbar}\, \Delta^- \mathrm{Im}(\psi_1^0 \pm \psi_1^3) - \frac{\beta}{\hbar}\,(\Delta^- \mathrm{Re}\psi_1^2 \pm \Delta^+ \mathrm{Im}\psi_1^1) \\ & \pm \frac{2\beta}{\hbar}\, \mathrm{Im}\psi_1^1 \pm \frac{\beta}{g\tau}\, \vartheta_1 \mathrm{Re}\psi_1^2 \pm \frac{\beta^2}{4g^2\tau}\, \vartheta_2\, \mathrm{Im}\varphi_2^3 \pm \frac{\beta^2}{2\hbar g}\, \Delta^- \mathrm{Im}\varphi_2^0. \end{aligned} \tag{A.8}$$

Inserting (A.3) and this expression into (4.8) and using (4.17) yield (4.20), (4.22), and (4.23). Up to order λ^2, we have thus proven the following statement:

By using the compatibility conditions in the hierarchy of (4.11), (4.12), *we obtain the same balance equations for* $n^\pm$ *as by direct substitution of the solutions of the hierarchy into* (3.25) *(which arise from integration of the kinetic equation over* k*).*

REFERENCES

[1] P. L. Bhatnagar, E. P. Gross, and M. Krook, *A model for collision processes in gases.* I. *Small amplitude processes in charged and neutral one-component systems*, Phys. Rev., 94 (1954), pp. 511–525.

[2] L. L. Bonilla and R. Escobedo, *Wigner-Poisson and nonlocal drift-diffusion model equations for semiconductor superlattices*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1253–1272.

[3] L. L. Bonilla and H. T. Grahn, *Nonlinear dynamics of semiconductor superlattices*, Rep. Progr. Phys., 68 (2005), pp. 577–683.

[4] L. L. Bonilla and F. J. Higuera, *The onset and end of the Gunn effect in extrinsic semiconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1625–1649.

[5] L. L. Bonilla, R. Escobedo, and A. Perales, *Generalized drift-diffusion model for miniband superlattices*, Phys. Rev. B, 68 (2003), article 241304(R).

[6] L. L. Bonilla, R. Escobedo, M. Carretero, and G. Platero, *Multiquantum well spin oscillator*, Appl. Phys. Lett., 91 (2007), article 092102.

[7] C. Cercignani, I. M. Gamba, and C. D. Levermore, *A drift-collision balance for a Boltzmann–Poisson system in bounded domains*, SIAM J. Appl. Math., 61 (2001), pp. 1932–1958.

[8] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Stat. Phys., 112 (2003), pp. 587–628.

[9] R. Escobedo and L. L. Bonilla, *Numerical methods for a quantum drift-diffusion equation in semiconductor physics*, J. Math. Chem., 40 (2006), pp. 3–13.

[10] H. Haug and A.-P. Jauho, *Quantum Kinetics in Transport and Optics of Semiconductors*, Springer, Berlin, 1996.

[11] A. A. Ignatov and V. I. Shashkin, *Bloch oscillations of electrons and instability of space-charge waves in semiconductor superlattices*, Soviet Phys. JETP, 66 (1987), pp. 526–530.

[12] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics*, W. A. Benjamin, New York, 1962.

[13] E. O. Kane, *The $k{\cdot}p$ method*, in Physics of III-V Compounds, Semiconductors and Semimetals, Vol. 1, R. Willardson and A. Beer, eds., Academic Press, New York, 1966, Chap. 3, pp. 75–100.

[14] P. Kleinert, V. V. Bryksin, and O. Bleibaum, *Spin accumulation in lateral semiconductor superlattices induced by a constant electric field*, Phys. Rev. B, 72 (2005), article 195311.

[15] X. L. Lei, *Distribution function and balance equations of drifting Bloch electrons in an electric field*, Phys. Rev. B, 51 (1995), pp. 5526–5530.

[16] X. L. Lei and C. S. Ting, *Theory of nonlinear electron transport for solids in a strong electric field*, Phys. Rev. B, 30 (1984), pp. 4809–4812.

[17] G. Platero and R. Aguado, *Photon-assisted transport in semiconductor nanostructures*, Phys. Rep., 395 (2004), pp. 1–157.

[18] E. I. Rashba, *Properties of semiconductors with an extremum loop.* 1. *Cyclotron and combinational resonance in a magnetic field perpendicular to the plane of the loop*, Sov. Phys. Solid State, 2 (1960), pp. 1224–1238.

[19] D. Sánchez, A. H. MacDonald, and G. Platero, *Field-domain spintronics in magnetic semiconductor multiple quantum wells*, Phys. Rev. B, 65 (2002), article 035301.

[20] E. Schomburg, T. Blomeier, K. Hofbeck, J. Grenzer, S. Brandl, I. Lingott, A. A. Ignatov, K. F. Renk, D. G. Pavelev, Y. Koschurinov, B. Y. Melzer, V. M. Ustinov, S. V. Ivanov, A. Zhukov, and P. S. Kopev, *Current oscillations in superlattices with different miniband widths*, Phys. Rev. B, 58 (1998), pp. 4035–4038.

[21] A. Wacker, *Semiconductor superlattices: A model system for nonlinear transport*, Phys. Rep., 357 (2002), pp. 1–111.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 514–523

THERMAL BLOW-UP IN A SUBDIFFUSIVE MEDIUM*

W. E. OLMSTEAD[†] AND CATHERINE A. ROBERTS[‡]

Abstract. The problem of thermal blow-up in a subdiffusive medium is examined within the framework of a fractional heat equation with a nonlinear source term. This model establishes that a thermal blow-up always occurs when a finite strip of subdiffusive material is exposed to the effects of a localized, high-energy source. This behavior is distinctly different from the classical diffusion case in which a blow-up can be avoided by locating the site of the energy source sufficiently close to one of the cold ends of the strip. The asymptotic growth of the solution near blow-up is determined for a nonlinear source whose output increases with temperature in either an algebraic or exponential manner. The blow-up growth rate is found to depend upon the anomalous diffusion parameter that defines the subdiffusive medium. This suggests that such media might be characterized by their response to a reaction-diffusion process.

Key words. subdiffusion, thermal blow-up, Volterra equation, asymptotic growth

AMS subject classifications. 35K60, 45D05, 80A20, 35B40

DOI. 10.1137/080714075

1. Introduction.

The problem of a thermal blow-up in a subdiffusive medium is examined. The diffusion of heat is retarded in materials with subdiffusive properties, thereby allowing the presence of a high-energy source to be considerably more effective in producing extreme temperature growth. The results here will demonstrate that a thermal blow-up will always occur, regardless of the proximity of a cold boundary.

The underlying physics of subdiffusion is associated with a medium in which the mean square displacement of Brownian motion evolves on a slower-than-normal time scale. That is,

$$\left\langle X^2 \right\rangle \sim C\, t^{\alpha}, \quad 0 < \alpha < 1, \tag{1}$$

where α is the anomalous diffusion parameter. The limiting case of $\alpha = 1$ corresponds to classical (Gaussian) diffusion. From the viewpoint of a random walk, a subdiffusive process exhibits an infinitely long average time for the occurrence of a finite jump, thereby implying a diminished capacity for the flux of thermal energy.

Subdiffusion occurs in a variety of applications as discussed in the review papers [4], [5], [11]. For the problem presented here, it is convenient to think of the application to certain porous materials in which microscopic pores are filled with a substance that has a lower conductivity than that of the basic matrix material as described in [1], [3]. A continuum model of a subdiffusive material is consistent with the scenario in which the pore size is small in comparison to $\sqrt{\langle X^2 \rangle}$.

The modeling of subdiffusive phenomena that obeys (1) has motivated the implementation of fractional differential operators as discussed in [4]. For the initial-boundary value problem considered here, a fractional diffusion equation with a localized, high-energy source will be defined for a finite strip of subdiffusive material.

*Received by the editors January 23, 2008; accepted for publication (in revised form) July 15, 2008; published electronically November 14, 2008.
http://www.siam.org/journals/siap/69-2/71407.html

[†]Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208-3125 (weo@northwestern.edu).

[‡]Mathematics and Computer Science, College of the Holy Cross, Worcester, MA 01610 (croberts@holycross.edu).

The ends of the strip are maintained at zero temperature so that some energy can be dissipated into the surroundings. In the context of a porous material, this model can be viewed as the thermal response associated with localized combustion of the porous material.

The investigation of a blow-up solution is carried out by converting the initial-boundary value problem to a nonlinear integral equation that governs the temperature at the site of the localized source. The resulting Volterra equation lends itself to the analytical techniques presented in [2], [6], [8], [9], and [10].

The results developed here for blow-up in a subdiffusive medium can be compared with those of [6] for the case of classical diffusion. In the classical diffusive problem, the results of [6] demonstrate that the occurrence of a blow-up with Dirichlet boundary conditions depends upon the proximity of the localized source to one of the ends of the strip. In particular, if the site of the source is located sufficiently close to either end, a blow-up will not occur. This implies that in the case of classical diffusion, the cold boundary can draw away enough heat from a nearby source to keep the temperature bounded throughout the strip.

In contrast to classical diffusion, the results here for the subdiffusive case will demonstrate that a blow-up will always occur, no matter how close the localized source is placed to a cold boundary. This behavior will be found to hold for all values of the anomalous diffusion parameter that correspond to the subdiffusive range. Further results developed here will show that the temporal growth of the temperature near blow-up can be characterized by the anomalous diffusion parameter.

2. Mathematical formulation. It is assumed that the temperature $T(x,t)$ in the strip of subdiffusive material satisfies the one-dimensional fractional diffusion equation given by

$$\frac{\partial T(x,t)}{\partial t} = \frac{\partial^2}{\partial x^2} D_t^{1-\alpha}[T(x,t)] + \delta(x-a)g[T(a,t)], \quad 0 < x < \ell, \quad t > 0, \tag{2}$$

$$T(0,t) = 0, \quad T(\ell,t) = 0, \quad t > 0, \tag{3}$$

$$T(x,0) = 0, \quad 0 \le x \le \ell. \tag{4}$$

The fractional derivative operator $D_t^{1-\alpha}$ is defined by

$$D_t^{1-\alpha}[T(x,t)] \equiv \frac{1}{\Gamma(\alpha)} \frac{\partial}{\partial t} \int_0^t \left(t-t'\right)^{\alpha-1} T\left(x,t'\right) \,\mathrm{d}t', \quad 0 < \alpha < 1, \tag{5}$$

where $\Gamma(\alpha)$ is the gamma function. This operator is introduced to model diffusive behavior that is consistent with (1). The particular form of (5) is known as the Riemann–Liouville fractional derivative. This form, as well as other alternative versions, is discussed in [7]. The limiting case of $\alpha = 1$ is associated with classical diffusion since D_t^0 is the identity operator.

The energy source term in (2) has been both localized and intensified by introducing the multiplicative delta function $\delta(x-a)$, $0 < a < \ell$. The nonlinearity $g(T)$ is assumed to have the properties

$$g(T) > 0, \quad g'(T) > 0, \quad g''(T) > 0, \quad T \ge 0, \tag{6}$$

which is typical for reaction-diffusion phenomena.

A physical interpretation of (2)–(6) is the temperature distribution in a strip of microscopically porous material that is capable of a chemical reaction only within a narrow zone centered at $x = a$. The source term in (2) includes a delta function to localize and intensify the reaction at $x = a$, while the nonlinearity $g(T)$ approximates an Arrhenius-type energy release.

3. Conversion to an integral equation. To investigate a possible blow-up solution of (2)–(6), it is advantageous to convert the initial-boundary value problem into an equivalent integral equation. This will be accomplished through the use of the Green's function $G_\alpha(x,t|\xi,0)$ that satisfies

$$\frac{\partial}{\partial t}G_\alpha(x,t|\xi,0) = \frac{\partial^2}{\partial x^2}D_t^{1-\alpha}\left[G_\alpha(x,t|\xi,0)\right] + \delta(x-\xi)\delta(t), \quad 0 < x < \ell, \quad t > 0^-, \tag{7}$$

$$G_\alpha(0,t|\xi,0) = 0, \quad G_\alpha(\ell,t|\xi,0) = 0, \quad t > 0, \tag{8}$$

$$G_\alpha(x,0^-|\xi,0) = 0, \quad 0 \le x \le \ell. \tag{9}$$

It follows from (2)–(4) and (7)–(9) that

$$T(x,t) = \int_0^t \int_0^\ell G_\alpha(x,t-s|\xi,0)\delta(\xi-a)g[T(a,s)]\,\mathrm{d}\xi\,\mathrm{d}s, \quad 0 \le x \le \ell, \quad t > 0. \tag{10}$$

Utilizing the sifting property of the delta function allows (10) to be reduced to

$$T(x,t) = \int_0^t G_\alpha(x,t-s|a,0)g[T(a,s)]\,\mathrm{d}s, \quad 0 \le x \le \ell, \quad t > 0. \tag{11}$$

It is clear from (11) that if $T(a,t)$ is known, then $T(x,t)$ is determined for $0 \le x \le \ell$, $t \ge 0$. Moreover, it is seen from (11) that any blow-up solution of (2)–(6) must be associated with a blow-up of $T(a,t)$.

In order to determine $T(a,t)$, set $x = a$ in (11), which produces the integral equation

$$u(t) = \int_0^t k(t-s)g[u(s)]\,\mathrm{d}s, \quad 0 \le t < \infty, \tag{12}$$

where

$$u(t) \equiv T(a,t) \tag{13}$$

and

$$k(t) \equiv G_\alpha(a,t|a,0). \tag{14}$$

Thus, the investigation of a possible blow-up solution of the initial-boundary value problem (2)–(6) has been reduced to the analysis of the integral equation (12).

To analyze (12), it is essential to know the properties of the kernel $k(t)$. Since those properties follow from $G_\alpha(x,t|\xi,0)$, it is necessary to solve (7)–(9). Results presented in [12] imply that the solution of (7)–(9) can be expressed in terms of the solution for the classical diffusion case in which $\alpha = 1$. That is,

$$G_\alpha(x,t|\xi,0) = \int_0^\infty f_\alpha(z)G_1(x,t^\alpha z|\xi,0)\,\mathrm{d}z. \tag{15}$$

To define $f_\alpha(z)$, it is necessary to introduce the definition of the Mellin transform

$$M\left[v(z);r\right] \equiv \int_0^\infty z^{r-1} v(z)\,\mathrm{d}z. \tag{16}$$

In [12], $f_\alpha(z)$ is introduced as an inverse Mellin transform as given by

$$f_\alpha(z) = M^{-1}\left[\frac{\Gamma(r)}{\Gamma(1-\alpha+\alpha r)}\right] = \sum_{j=0}^{\infty} \frac{(-1)^j z^j}{j!\,\Gamma(1-\alpha-\alpha j)}, \quad z \geq 0. \tag{17}$$

It follows from (17) that

$$f_\alpha(z) \geq 0, \quad z \geq 0, \tag{18}$$

and

$$f_\alpha(z) \to 0 \text{ exponentially as } z \to \infty. \tag{19}$$

For the classical diffusion case in which $\alpha = 1$, the solution of (7)–(9) can be expressed either as a Fourier sine series expansion,

$$G_1(x,t|\xi,0) = \frac{2H(t)}{\ell} \sum_{n=1}^{\infty} \sin\frac{n\pi\xi}{\ell}\,\sin\frac{n\pi x}{\ell}\,\exp\left(-\frac{n^2\pi^2}{\ell^2}t\right), \tag{20}$$

or as an image expansion,

$$G_1(x,t|\xi,0) = \frac{H(t)}{2(\pi t)^{\frac{1}{2}}} \sum_{n=-\infty}^{\infty} \left\{\exp\left[\frac{(x-\xi-2n\ell)^2}{-4t}\right] - \exp\left[\frac{(x+\xi-2n\ell)^2}{-4t}\right]\right\}, \tag{21}$$

where $H(t)$ is the Heaviside function.

Two versions of the solution to (7)–(9) can be derived from (15), (20), and (21). From (15) and (20) it follows that

$$G_\alpha(x,t|\xi,0) = \frac{2H(t)}{\ell} \sum_{n=1}^{\infty} \sin\frac{n\pi\xi}{\ell}\,\sin\frac{n\pi x}{\ell} \int_0^\infty f_\alpha(z)\exp\left(-\frac{n^2\pi^2}{\ell^2}t^\alpha z\right)\mathrm{d}z, \tag{22}$$

while from (15) and (21) it follows that

$$\begin{aligned} G_\alpha(x,t|\xi,0) = \frac{H(t)}{2\pi^{\frac{1}{2}}t^{\frac{\alpha}{2}}} \sum_{n=-\infty}^{\infty} \int_0^\infty & z^{-\frac{1}{2}} f_\alpha(z) \\ & \times \left\{\exp\left[\frac{(x-\xi-2n\ell)^2}{-4t^\alpha z}\right] - \exp\left[\frac{(x+\xi-2n\ell)^2}{-4t^\alpha z}\right]\right\}\mathrm{d}z. \end{aligned} \tag{23}$$

Two alternate expressions for the kernel $k(t)$, as defined by (14), follow from (22) and (23) either as

$$k(t) = \frac{2}{\ell}\sum_{n=1}^{\infty} \sin^2\frac{n\pi a}{\ell} \int_0^\infty f_\alpha(z)\exp\left(-\frac{n^2\pi^2}{\ell^2}t^\alpha z\right)\mathrm{d}z \tag{24}$$

or as

$$k(t) = \frac{1}{2\pi^{\frac{1}{2}}t^{\frac{\alpha}{2}}} \sum_{n=-\infty}^{\infty} \int_0^\infty z^{-\frac{1}{2}} f_\alpha(z)\left\{\exp\left[-\frac{n^2\ell^2}{t^\alpha z}\right] - \exp\left[-\frac{(a-n\ell)^2}{t^\alpha z}\right]\right\}\mathrm{d}z. \tag{25}$$

In (24)–(25), $H(t)$ has been dropped since it is superfluous to the interpretation of $k(t-s)$ in (12).

Various properties of $k(t)$ can be derived from (24)–(25). It is easily seen that $k(t)$ is a continuously differentiable function for $0<t<\infty$, and

$$k(t)>0, \quad k'(t)<0, \quad 0<t<\infty. \tag{26}$$

The asymptotic behavior of $k(t)$ as $t\to 0$ and as $t\to\infty$ is important in the analysis of (12). As $t\to 0$, the integrals in (25) are all negligible compared to the $n=0$ term. It then follows that

$$k(t)\sim \frac{1}{2\pi^{\frac{1}{2}}t^{\frac{\alpha}{2}}}\int_0^\infty z^{-\frac{1}{2}}f_\alpha(z)\,\mathrm{d}z=\frac{1}{2\,\Gamma\left(1-\frac{\alpha}{2}\right)t^{\frac{\alpha}{2}}} \text{ as } t\to 0. \tag{27}$$

As $t\to\infty$, it is useful to rescale the integration variable in (24) to obtain

$$k(t)=\frac{2}{\ell t^\alpha}\sum_{n=1}^{\infty}\sin^2\frac{n\pi a}{\ell}\int_0^\infty f_\alpha(z/t^\alpha)\exp\left(-\frac{n^2\pi^2}{\ell^2}z\right)\mathrm{d}z, \tag{28}$$

from which follows

$$k(t)\sim\left(\frac{2\ell}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin^2\frac{n\pi a}{\ell}\right)\frac{1}{\Gamma(1-\alpha)\,t^\alpha}=\frac{a(\ell-a)}{\ell\,\Gamma(1-\alpha)\,t^\alpha} \text{ as } t\to\infty. \tag{29}$$

4. Blow-up solution. A physical interpretation of a blow-up solution to (12), and hence to (2)–(6), is that the subdiffusive medium is unable to conduct enough heat away from the energy source to prevent a thermal runaway. A measure of the subdiffusive medium's ability to conduct heat is given by $I(t)$, which is defined by

$$I(t)\equiv\int_0^t k(s)\,\mathrm{d}s, \quad t\ge 0. \tag{30}$$

A measure of the strength of the energy source is given by κ, which is defined by

$$\kappa\equiv\int_0^\infty\frac{\mathrm{d}u}{g(u)}<\infty, \tag{31}$$

with the assumption that the integral is finite. Another measure of the strength of the energy source is given by Λ, where

$$\Lambda\equiv\sup_{0\le u<\infty}\left[\frac{u}{g(u)}\right]. \tag{32}$$

The properties of $g(u)$ in (6) ensure that

$$\Lambda\le\kappa<\infty. \tag{33}$$

The essential results on existence and blow-up of a solution to (12) are derived in [2], [6], [10]. The basic results on existence, uniqueness, and blow-up are given by the following lemmas.

Lemma 1. *Let $k(t)\ge 0$ be continuous for $0<t<\infty$ and integrable as $t\to 0$. Then (12) has a unique continuous solution for $0\le t<t^*$, where $t^*<\infty$ if t^* is such that*

$$I\left(t^*\right)=\Lambda, \tag{34}$$

while $t^* = \infty$ *if*

$$I(t) < \Lambda, \quad 0 \le t < \infty. \tag{35}$$

Lemma 2. *Let* $k(t) \ge 0$ *be continuous and nonincreasing for* $0 < t < \infty$ *and integrable as* $t \to 0$. *Then whenever there exists a* $t^{**} < \infty$ *such that*

$$I\left(t^{**}\right) = \kappa, \tag{36}$$

it follows that (12) *cannot have a continuous solution for* $t \ge t^{**}$.

The nonexistence of a global solution to (12) is associated with the blow-up behavior

$$u(t) \to \infty \text{ as } t \to \hat{t} < \infty. \tag{37}$$

An implication of Lemmas 1 and 2 is that when (12) has a blow-up solution, the blow-up time $\hat{t}$ can be bounded as

$$0 < t^* \le \hat{t} \le t^{**} < \infty, \tag{38}$$

where t^* and t^{**} are determined by (34) and (36), respectively.

In order to apply Lemmas 1 and 2, it is essential to know the properties of $I(t)$. From the properties of $k(t)$ expressed by (26), (27), and (29), it follows that $I(t)$ is continuous for $0 \le t < \infty$ and

$$I(t) > 0, \quad I'(t) > 0, \quad 0 < t < \infty, \quad I(0) = 0, \tag{39}$$

with the asymptotic behavior

$$I(t) \sim \frac{1}{2\left(1 - \frac{\alpha}{2}\right)\Gamma\left(1 - \frac{\alpha}{2}\right)} t^{1-\frac{\alpha}{2}} \text{ as } t \to 0, \tag{40}$$

and

$$I(t) \sim \frac{a(\ell - a)}{\ell\,\Gamma(2 - \alpha)} t^{1-\alpha} \text{ as } t \to \infty. \tag{41}$$

In view of the behavior of $I(t)$ provided by (39)–(40), it is clear that (34) will be satisfied by some t^*, $0 < t^* < \infty$. Moreover, the asymptotic growth of $I(t)$ provided by (41) ensures that (36) will be satisfied by some $t^{**} < \infty$. Thus we obtain the following theorem.

Theorem 3. *The integral equation* (12) *has a unique, continuous solution for* $0 \le t < t^* < \infty$. *That solution ultimately becomes unbounded as* $t \to \hat{t} < \infty$, *where* $\hat{t} > 0$ *satisfies* (38).

It is important to note that the result of Theorem 3 relies upon the restrictions that $0 < a < \ell$ and $0 < \alpha < 1$. Since (36) must be satisfied by some sufficiently large but finite value of t^{**} for the subdiffusion problem, a blow-up will ultimately occur no matter how close the energy source is located to one of the cold endpoints of the strip. This is contrary to the case of classical diffusion. For $\alpha = 1$, it was demonstrated in [6] that (35) could always be satisfied by making $a(\ell - a)$ sufficiently small, thereby ensuring a unique continuous solution of (12) for all $t \ge 0$. This result is also implied by (41) in the limit $\alpha \to 1$, since $I(t)$ remains bounded as $t \to \infty$.

5. Blow-up growth rate. The growth rate of $u(t)$ near blow-up can be determined from an asymptotic analysis of (12) as $t \to \hat{t}$. To carry out that analysis, it is appropriate to follow the approach developed in [9] for a class of nonlinear Volterra equations that includes (12). As shown in [9], the blow-up growth rate is determined by the asymptotic behavior of $k(t)$ as $t \to 0$ and the asymptotic behavior of $g(u)$ as $u \to \infty$.

The asymptotic behavior of $k(t)$ as $t \to 0$ is given by (27), which indicates its dependence on the anomalous diffusion parameter α. As for the asymptotic behavior of $g(u)$ near blow-up, the results here will be confined to the special cases in which $g(u)$ has either (i) algebraic growth or (ii) exponential growth as $u \to \infty$.

For the asymptotic analysis of (12), it is convenient to introduce the changes of variables

$$\eta = \left(\hat{t} - t\right)^{-1} - \eta_0, \quad \eta_0 = \hat{t}^{-1}, \quad w(\eta) = u(t). \tag{42}$$

This transformation converts (12) to the form

$$w(\eta) = \int_0^\eta k\left\{(\eta - \eta')\left[(\eta' + \eta_0)(\eta + \eta_0)\right]^{-1}\right\}(\eta' + \eta_0)^{-2} g[w(\eta')]\,\mathrm{d}\eta', \quad 0 \le \eta < \infty. \tag{43}$$

In terms of the new variables, the blow-up defined by (37) is expressed as

$$w(\eta) \to \infty \text{ as } \eta \to \infty. \tag{44}$$

Following the methods of [9], let $\eta' = \eta\tau$ so that (5) becomes

$$w(\eta) = \eta\, Q(\eta), \quad 0 \le \eta < \infty, \tag{45}$$

where

$$Q(\eta) = \int_0^1 k\left\{\eta(1-\tau)[(\eta\tau + \eta_0)(\eta + \eta_0)]^{-1}\right\}(\eta\tau + \eta_0)^{-2} g[w(\eta\tau)]\,\mathrm{d}\tau. \tag{46}$$

Thus, the blow-up growth rate of $w(\eta)$ can be determined from an asymptotic analysis of (45) as $\eta \to \infty$. It is shown in [9] that the leading order behavior of $Q(\eta)$ as $\eta \to \infty$ is determined by the leading order behavior of $k(t)$ as $t \to 0$. It then follows from (27) that

$$Q(\eta) \sim \frac{1}{2\,\Gamma\left(1 - \frac{\alpha}{2}\right)} \int_0^\infty (1-\tau)^{-\frac{\alpha}{2}} H(1-\tau)\Psi(\eta\tau)\,\mathrm{d}\tau \text{ as } \eta \to \infty, \tag{47}$$

where $H(\tau)$ is the Heaviside function and

$$\Psi(\eta\tau) = (\eta\tau + \eta_0)^{-2+\frac{\alpha}{2}} g[w(\eta\tau)]. \tag{48}$$

Following the method of [9], the integral in (47) is converted to an integral in the complex z-plane by the application of the Parseval formula for Mellin transforms. This gives

$$Q(\eta) \sim \frac{1}{4\pi i\,\Gamma\left(1 - \frac{\alpha}{2}\right)} \int_{c-i\infty}^{c+i\infty} M\left[(1-\tau)^{-\frac{\alpha}{2}} H(1-\tau); 1-r\right] M[\Psi(\eta\tau); r]\,\mathrm{d}r. \tag{49}$$

The Mellin transforms in (49) are consistent with the definition provided in (16). Further simplification of (49) is achieved by noting that

$$M\left[(1-\tau)^{-\frac{\alpha}{2}}H(1-\tau);1-r\right]=\frac{\Gamma\left(1-\frac{\alpha}{2}\right)\Gamma(1-r)}{\Gamma\left(2-\frac{\alpha}{2}-r\right)} \tag{50}$$

and

$$M[(\Psi(\eta\tau);r]=\eta^{-r}M[\Psi(\tau);r]. \tag{51}$$

This allows the integral equation (45) to be replaced by the asymptotic equation

$$w(\eta)\sim\frac{1}{4\pi i}\int_{c-i\infty}^{c+i\infty}\eta^{1-r}\frac{\Gamma(1-r)}{\Gamma\left(2-\frac{\alpha}{2}-r\right)}M[\Psi(\tau);r]\,\mathrm{d}r\text{ as }\eta\to\infty. \tag{52}$$

To proceed with the asymptotic analysis, it is necessary to introduce an assumption about the growth of $g(u)$. Consider the case in which $g(u)$ has algebraic growth:

$$g(u)\sim u^{m}(\eta),\ m>1\text{ as }\eta\to\infty. \tag{53}$$

To obtain an asymptotic solution of (52) for this case, it is assumed that

$$u(\eta)\sim A\eta^{p},\ p>0\text{ as }\eta\to\infty. \tag{54}$$

The constants A and p are to be determined by satisfying (52) to leading order.

From (48), (53), and (54), it follows that

$$\Psi(\eta)\sim A^{m}\eta^{-2+\frac{\alpha}{2}+mp}\text{ as }\eta\to\infty. \tag{55}$$

By imposing the restriction that $1>2-\frac{\alpha}{2}-mp$, it follows that $M[\Psi;r]$ has a simple pole at $r=2-\frac{\alpha}{2}-mp<1$ and

$$M[\Psi;r]\sim-\frac{A^{m}}{r-\left(2-\frac{\alpha}{2}-mp\right)}\text{ as }r\to2-\frac{\alpha}{2}-mp. \tag{56}$$

Now the leading asymptotic contribution from the integral in (52) comes from the pole implied by (56). As the vertical path of integration is displaced to the right, that pole is encountered before the pole at $r=1$ arising from $\Gamma(1-r)\sim-(r-1)^{-1}$ as $r\to1$. Thus (52) takes the form

$$A\,\eta^{p}\sim\frac{A^{m}\,\Gamma\left(mp+\frac{\alpha}{2}-1\right)}{2\,\Gamma(mp)}\,\eta^{mp+\frac{\alpha}{2}-1}\text{ as }\eta\to\infty. \tag{57}$$

From (57) it is concluded that

$$p=\frac{1-\frac{\alpha}{2}}{m-1},\quad A=\left\{\frac{2\,\Gamma\left[\frac{m\left(1-\frac{\alpha}{2}\right)}{m-1}\right]}{\Gamma\left[\frac{\left(1-\frac{\alpha}{2}\right)}{m-1}\right]}\right\}^{\frac{1}{m-1}}. \tag{58}$$

These results are seen to be consistent with the original constraint that $1>2-\frac{\alpha}{2}-mp=1-\left[\left(1-\frac{\alpha}{2}\right)/(m-1)\right]$. The complement of this constraint leads to a contradiction of any leading order asymptotic match in (52).

In view of (54) and (58), the asymptotic growth of the solution to (12) near blow-up is given by

$$u(t) \sim A(\hat{t}-t)^{-p} \text{ as } t \to \hat{t} \tag{59}$$

for the case in which $g(u)$ grows algebraically as specified by (53).

Next consider the case in which $g(u)$ has exponential growth,

$$g(u) \sim e^{u} \text{ as } \eta \to \infty. \tag{60}$$

To obtain an asymptotic solution of (52) for this case, it is assumed that

$$u(\eta) \sim \log\left(A\eta^{p}\right) \sim p\log\eta \text{ as } \eta \to \infty. \tag{61}$$

The constants A and p are to be determined by satisfying (52) to leading order.

From (48), (60), and (61), it follows that

$$\Psi(\eta) \sim A\eta^{-2+\frac{\alpha}{2}+p} \text{ as } \eta \to \infty. \tag{62}$$

It follows that $M[\Psi; r]$ has a simple pole at $r = 2 - \frac{\alpha}{2} - p$ and

$$M[\Psi; r] \sim -\frac{A}{r-\left(2-\frac{\alpha}{2}-p\right)} \text{ as } r \to 2-\frac{\alpha}{2}-p. \tag{63}$$

In order for the leading asymptotic contribution from the integral in (52) to yield a logarithmic term that will match (61), it is necessary that the simple pole for $M[\Psi; r]$ coalesce with that arising from $\Gamma(1-r) \sim -(r-1)^{-1}$ as $r \to 1$. This requires that

$$p = 1 - \frac{\alpha}{2}. \tag{64}$$

As the vertical path of integration is displaced to the right, the leading order contribution from the double pole reduces (52) to

$$\left(1-\frac{\alpha}{2}\right)\log\eta \sim \frac{A}{2\,\Gamma\left(1-\frac{\alpha}{2}\right)}\log\eta \text{ as } \eta \to \infty. \tag{65}$$

An asymptotic match in (65) is achieved by taking

$$A = 2\left(1-\frac{\alpha}{2}\right)\Gamma\left(1-\frac{\alpha}{2}\right) = 2\,\Gamma\left(2-\frac{\alpha}{2}\right), \tag{66}$$

although this constant plays no role in the leading order behavior. In view of (61) and (64), the leading order asymptotic growth of the solution to (12) near blow-up is given by

$$u(t) \sim \left(1-\frac{\alpha}{2}\right)\log\left(\frac{1}{\hat{t}-t}\right) \text{ as } t \to \hat{t} \tag{67}$$

for the case in which $g(u)$ grows exponentially as specified by (60).

6. Conclusions. For the case of a finite strip of subdiffusive material subjected to a localized high-energy source, as modeled by (2)–(6), a thermal blow-up always occurs. Unlike the case of classical (Gaussian) diffusion, the blow-up cannot be averted by locating the site of the source sufficiently close to a cold boundary. This result is consistent with the physics of subdiffusion in which the flux of energy is retarded.

The asymptotic growth of the temperature near blow-up was derived for nonlinear energy sources that increase with temperature in either an (i) algebraic or (ii) exponential manner. In each case, the explicit dependence of the growth rate on the anomalous diffusion parameter was found. This suggests the possibility of experimentally determining the anomalous diffusion parameter from data collected during an appropriate reaction-diffusion process.

REFERENCES

[1] I. Ardelean, G. Farrher, and R. Kimmich, *Effective diffusion in partially filled nanoscopic and microscopic pores*, J. Optoelectron. Adv. Mater., 9 (2007), pp. 655–660.

[2] C. M. Kirk and W. E. Olmstead, *Blow-up in a reactive-diffusive medium with a moving heat source*, Z. Angew. Math. Phys., 53 (2002), pp. 147–159.

[3] Y. Li, G. Farrher, and R. Kimmich, *Sub- and superdiffusion molecular displacement laws in disordered porous media probed by nuclear magnetic resonance*, Phys. Rev. E, 74 (2006), article 066309.

[4] R. Metzler and J. Klafter, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.

[5] R. Metzler and J. Klafter, *The restaurant at the end of the random walk: Recent developments in the description of anomalous transport by fractional dynamics*, J. Phys. A, 37 (2004), pp. 161–208.

[6] W. E. Olmstead and C. A. Roberts, *Explosion in a diffusive strip due to a concentrated nonlinear source*, Methods Appl. Anal., 1 (1994), pp. 434–445.

[7] I. Podlubny, *Fractional Differential Equations*, Academic Press, New York, 1999.

[8] C. A. Roberts, *Recent results on blow-up and quenching for nonlinear Volterra equations*, J. Comput. Appl. Math., 205 (2007), pp. 736–743.

[9] C. A. Roberts and W. E. Olmstead, *Growth rates for blow-up solutions of nonlinear Volterra equations*, Quart. Appl. Math., 54 (1996), pp. 153–159.

[10] C. A. Roberts, D. G. Lasseigne, and W. E. Olmstead, *Volterra equations which model explosion in a diffusive medium*, J. Integral Equations Appl., 5 (1993), pp. 531–546.

[11] J. Trujillo, *Fractional models: Sub and super-diffusives, and undifferentiable solutions*, in Innovation in Engineering Computational Technology, B. H. V. Topping, G. Montero, and R. Montenegro, eds., Sax-Coburg Publ., 2006, pp. 371–402.

[12] M. M. Wyss and W. Wyss, *Evolution, its fractional extension and generalization*, Fract. Calc. Appl. Anal., 4 (2001), pp. 273–284.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 524–551

STABILIZING ROLE OF A CURVATURE CORRECTION TO LINE TENSION*

RICCARDO ROSSO[†] AND MARCO VERANI[‡]

Abstract. We study the effects that a curvature correction to the line tension has on the equilibrium and stability of liquid droplets laid upon a rigid substrate. In the simple case of cylindric liquid bridges we prove that even a tiny curvature correction prevents the onset of wildly oscillating perturbations that would make the contact line unstable if a negative line tension were present alone. However, if the curvature correction is not large enough, unstable modes that are not related to the classical Rayleigh instability can persist.

Key words. line tension, wetting, curvature corrections

AMS subject classification. 76B45

DOI. 10.1137/070710585

1. Introduction.

Since Gibbs' fundamental paper [1] on the equilibrium of heterogeneous substances, there has been an increasing interest in modeling both the statics and the dynamics of multiphase bodies. In particular, a faithful description of the interface separating two different phases has been sought along different lines originating from either the continuum point of view or the microscopic point of view that relies upon statistical mechanics. In his original approach, Gibbs modeled the thin interfacial *three*-dimensional region where the physical properties of two adjoining phases rapidly change as a *two*-dimensional surface, called the *dividing* surface, that separates two bulk regions where the phases are *homogeneous*. In general, extensive properties like energy or entropy differ in the real system and in the idealized one. Gibbs ascribed the excess energy or entropy to the dividing surface, adding surface energy and entropy to the bulk terms characterizing the homogeneous phases. The simplest surface energy introduced by Gibbs is proportional to the area of the dividing surface; the constant of proportionality—called the *surface tension*—being positive for stability reasons. Gibbs clearly stated that the surface energy he envisaged was appropriate only for *flat* or weakly curved interfaces, while in the general case other contributions depending on the interface curvature should enter the energy balance. It was Tolman [2] who first analyzed curvature corrections by expanding the surface tension pertaining to a spherical interface of radius R in powers of $1/R$. The length scale at which this correction is relevant is the *Tolman length* δ_T that has been found to be a molecular length both in numerical simulations [3] of Lennard–Jones fluids and in the analytic treatment of [4].

A general format to incorporate curvature corrections in the surface energy was sketched by Gibbs himself and later exploited, for instance, in [5], where the following

*Received by the editors December 11, 2007; accepted for publication (in revised form) June 24, 2008; published electronically December 3, 2008.
http://www.siam.org/journals/siap/69-2/71058.html

[†] Dipartimento di Matematica and CNISM, Università di Pavia, Via Ferrata 1, I-27100 Pavia, Italy (riccardo.rosso@unipv.it).

[‡] Mox-Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, I-20122 Milano, Italy (marco.verani@polimi.it).

expression for the surface tension

$$\gamma = \gamma_0 + \kappa\left(c_0 H + \frac{1}{2}H^2\right) + \hat{\kappa}K \tag{1.1}$$

was proposed, in which γ_0 is the surface tension for a flat interface, c_0 is the *spontaneous curvature* of the interface, and H and K are the *total* and the *Gaussian* curvatures of the interface, respectively, while the constitutive parameters κ and $\hat{\kappa}$ are *bending rigidities.* We record here that the Tolman length can be expressed as [5]

$$\delta_T = \frac{\kappa c_0}{\gamma_0}\,. \tag{1.2}$$

Equation (1.1) is a truncated expansion that also covers the case of nonspherical interfaces. It can also be noticed that the correction (1.1) to the surface tension transforms the surface energy into the Canham–Helfrich Hamiltonian so successfully employed in modeling biological membranes. To obtain more tractable expressions for the curvature corrections, an alternative procedure was recently put forward in [6] by performing a curvature expansion of the lowest order equation in the Born–Green–Yvon hierarchy.

Up to this point we have considered interfaces separating two distinct phases. However, *contact lines* where *three* different phases coexist at equilibrium also occur. A line energy proportional to the length of the contact curve had been introduced by Gibbs himself in [1] to model the excess free energy residing there. The constant of proportionality is called the *line tension.* Since line tension effects on equilibrium are detectable for systems in the submicron regime, its role had been neglected until experimental techniques became available, which allow explorations of these small-sized droplets. As a consequence, the impact of line tension on the equilibrium [7, 8] and the stability [9, 10, 11, 12, 13, 14, 15, 16, 17] of sessile droplets was studied thoroughly during the past decade. In particular, a controversy arose on the admissibility of a negative line tension within a continuum model. At variance with surface tension, Gibbs did not put restrictions on the sign of line tension, but it was proved in [9] that negative values of line tension would make the free-energy functional unbounded from below, and so make any equilibrium configuration unstable. Precisely, if the contact line is corrugated enough, a droplet at equilibrium can follow a path along which its energy is reduced. It was pointed out [12, 17, 18], however, that the characteristic wavelength induced by destabilizing perturbations on the equilibrium droplet could be a molecular length, detectable at a length scale outside the realm of a continuum model. In [18, 19] a criterion of *marginal stability* was proposed to estimate, roughly speaking, the number of stable modes for a given equilibrium configuration and for a given negative line tension. In this way, we could ascertain that negative line tensions as those reported in [20] were compatible with a large set of stable modes, and that the onset of instability was related to perturbations with so short a wavelength that they presumably operate at a scale where also curvature corrections to the line tension should be accounted for [18]. Incorporating these corrections into a continuum model to study their impact on the stability of sessile droplets is the aim of this paper.

In fact Boruvka and Neumann introduced long ago [21] curvature corrections for both the surface and the line energy, by building a formal theory where the free energy contains contributions depending on both the normal and the geodesic curvatures as well as on the geodesic torsion of the contact line, conceived as a curve lying either on the substrate or on the free surface of the liquid droplet. Here we do not insist in making all these differential-geometric properties enter the free-energy functional, as this would lead to a large number of constitutive parameters which, in turn, would make predictions rather difficult, if at all possible. So, we simply imagine that the line tension depends on the curvature σ of the contact line. In this sense our approach departs from that of Boruvka and Neumann, who did not consider corrections depending only on the curvature σ of the contact line since σ has no relation with either the free surface of the droplet or the substrate. However, on computing the first and the second variation of the line free energy, we will see that it is natural to consider deformations of a sessile droplet that map contact lines into contact lines. In this way, the geometry of the substrate is naturally coupled with that of the contact line and both the first and the second variation of the line energy depend on the geometric properties of the contact line, conceived as a curve on the substrate.

It should also be recalled that recent studies [22] have focused on the dependence of line tension upon the radius of curvature of the dividing line. We also mention that a different kind of curvature correction to line tension was studied in [23], where the dependence of line tension on the substrate's curvature was examined within an effective interfacial Hamiltonian approach, in the limit of weakly curved cylindric substrates.

The reader might wonder why we do not treat line and surface tension on the same footing, by assuming a dependence of the latter on curvature too. While in the next section we will give a technical reason for neglecting such corrections, a simple argument can be given, by comparing the typical energy of the term $\kappa c_0 \int_{\mathcal{S}} H \mathrm{d}A$ associated with Tolman's correction with the energy $\beta \int_{\mathcal{C}} \sigma^2 \mathrm{d}\ell$ associated with the curvature correction of line tension. Taking a spherical capsule of radius R, and recalling (1.2), the contribution due to curvature correction of line tension prevails whenever

$$R \ll \sqrt{\frac{\beta}{\gamma_0 \delta_T}}\,.$$

Since δ_T is a molecular length, the set of values of the ratio β/γ_0 that make this inequality obeyed by micron-sized droplets is nonempty.

This paper is organized as follows. In section 2 we introduce the curvature correction to line tension and we discuss the length scales hidden in our model. In section 3 we compute the first variation of the curvature correction arriving at a modified Young equation obeyed along the contact line. Here we also write down the second variation of the curvature-dependent correction, deferring to an appendix the lengthly calculations needed to obtain it. As an application, in section 4 we address the stability of a liquid bridge lying on a flat substrate that was explored without curvature correction in [16] and [17]. We prove that the curvature correction cancels the systematic instability induced by negative line tension for modes with arbitrarily short wavelengths, regardless of the magnitude of the correction. However, different stability scenarios can be singled out, depending on the magnitude of both the bare line tension and its curvature correction. The paper is closed by a section where we summarize our results and we outline some possible applications.

2. Free energy. We consider a sessile droplet $\mathcal{B}$ consisting of incompressible fluid (see Figure 1). Its boundary $\partial\mathcal{B}$ is naturally split as $\partial\mathcal{B} = \mathcal{S}^* \cup \mathcal{S}_*$, where the *adhering* surface $\mathcal{S}_*$ is laid on a rigid substrate. The portion $\mathcal{S}^*$ of $\partial\mathcal{B}$ that is not in

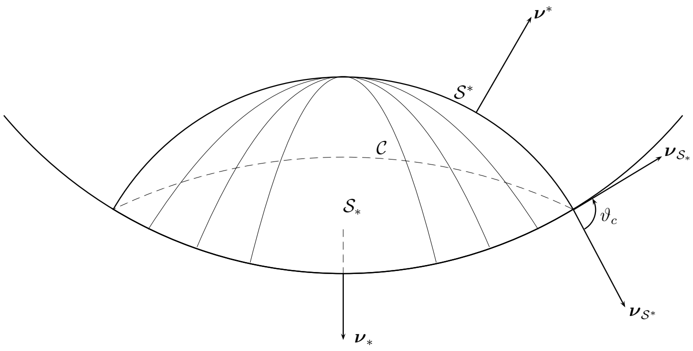


FIG. 1. *A sessile droplet laid on a rigid substrate. The boundary of the droplet is split into a free surface $\mathcal{S}^*$ and an adhering surface $\mathcal{S}_*$: on the former, the droplet is in contact with a vapor phase; on the latter it is in contact with the substrate. These surfaces meet along the contact line $\mathcal{C}$. The outer unit normal vectors $\boldsymbol{\nu}^*$ and $\boldsymbol{\nu}_*$ to $\mathcal{S}^*$ and $\mathcal{S}_*$ are also shown, together with the conormal unit vectors $\boldsymbol{\nu}_{\mathcal{S}^*}$ and $\boldsymbol{\nu}_{\mathcal{S}_*}$ to the contact line, conceived as a curve on $\mathcal{S}^*$ and $\mathcal{S}_*$, respectively. Finally, the contact angle ϑ_c, defined as the angle between $\boldsymbol{\nu}_{\mathcal{S}^*}$ and $\boldsymbol{\nu}_{\mathcal{S}_*}$, is also shown.*

contact with the substrate is referred to as the *free* surface. The curve $\mathcal{C} := \mathcal{S}^* \cap \mathcal{S}_*$ is the *contact* line, where three phases coexist at equilibrium. We shall assume, for simplicity, that $\mathcal{C}$ is connected. The equilibrium shapes of a droplet are the critical points of the free-energy functional

$$\mathcal{F}[\mathcal{B}] := \gamma_0 \int_{\mathcal{S}^*} \mathrm{d}a - w \int_{\mathcal{S}_*} \mathrm{d}a + \tau_0 \int_{\mathcal{C}} \mathrm{d}\ell + \beta \int_{\mathcal{C}} \sigma^2 \mathrm{d}\ell\,, \tag{2.1}$$

subject to the incompressibility constraint

$$\mathrm{vol}(\mathcal{B}) = \int_{\mathcal{B}} \mathrm{d}V = \text{constant.} \tag{2.2}$$

The functional (2.1) contains several contributions. First, $\gamma_0 \int_{\mathcal{S}^*} \mathrm{d}a$ accounts for the surface tension $\gamma_0 := \gamma_{lv} > 0$ associated with the interface between the liquid and the vapor phase. Here a is the area-measure on either $\mathcal{S}^*$ or $\mathcal{S}_*$. The term $-w \int_{\mathcal{S}_*} \mathrm{d}a$ is responsible for the excess energy at the solid-liquid interface. We introduced the *adhesion potential* $w > 0$ that is often expressed as $w = \gamma_{lv} - \gamma_{ls} + \gamma_{sv}$, that is, in terms of the surface tensions γ_{ls} and γ_{sv} associated with the liquid-solid and with the solid-vapor interfaces. We then consider two line energy contributions: the former, $\tau_0 \int_{\mathcal{C}} \mathrm{d}\ell$, models the *bare* line tension τ_0, a constant associated with an ideal straight contact line. The latter is $\beta \int_{\mathcal{C}} \sigma^2 \mathrm{d}\ell$, where σ is the curvature of the contact line $\mathcal{C}$, $\beta > 0$ is a constant parameter, and ℓ is the length-measure along $\mathcal{C}$. This term measures the curvature correction to the bare line tension and a squared dependence upon σ has been chosen to parallel the curvature correction to surface tension contained in (1.1): a linear term $\int_{\mathcal{C}} \sigma \mathrm{d}\ell$ is discarded as it would simply contribute a constant to $\mathcal{F}$, since $\mathcal{C}$ is a closed curve. If, for a moment, we consider the contact curve alone, leaving the droplet out of consideration, we are modelling $\mathcal{C}$ as an Euler's elastic curve. Pursuing this analogy, we could interpret β as a torsional rigidity of the contact line. In studying the relevance of curvature corrections on the stability of liquid droplets, Tolman length plays an ancillary role. In fact, we learned in previous work on this topic [16, 18] that the stability is determined by the natural boundary condition along the contact line that arises in the minimization of the second variation of the free-energy functional. Incorporating a curvature correction to the surface tension would add to this boundary condition terms depending on the curvature of the contact line that are qualitatively equivalent to those considered here. Moreover, this dependence would lead to a nonconstant mean curvature in the free surface profile that would render the normal mode analysis more intricate to follow. Hence, we think that the essential effects of curvature corrections are captured by just taking a curvature-dependent line tension. In the same vein, we did not consider bulk terms in (2.1) so that both gravity and a diluted interaction between the substrate and the droplet (see, e.g., [24]) are disregarded. It is true that, since bulk terms modify the equilibrium profile of the free surface, they could influence the stability analysis. However, the instability we are concerned with is driven by line tension and it seems plausible that bulk terms do not modify the qualitative features of the stability analysis.

We aim at exploring the stabilizing role of a curvature correction to the line tension and so we will assume hereafter $\tau_0 < 0$ since negative line tensions play a systematic, destabilizing effect. As we discussed in several geometries [16, 18, 19], conditionally stable equilibria in the presence of negative line tension are possible provided that $|\tau_0|$ is sufficiently small. In this case, it can be shown that the typical wavelength of destabilizing modes is a molecular length that lies outside the realm of application of the continuum picture adopted here. We expect that a curvature correction penalizing wild oscillations of the contact line could enhance the stability of an equilibrium configuration, even if line tension is negative.

We now digress slightly to introduce the characteristic lengths hidden in our model. A first length scale ℓ_τ can be defined as the typical linear dimension of a droplet for which the surface energy and the line energy associated with bare line tension τ_0 have the same order of magnitude:

$$\gamma_0 \int_{\mathcal{S}^*} \mathrm{d}a \approx |\tau_0| \int_{\mathcal{C}} \mathrm{d}\ell \,.$$

If $\mathcal{S}^*$ is a spherical capsule of radius ℓ_τ so that $\mathcal{C}$ is a circumference of radius $R \propto \ell_\tau$, we obtain $\ell_\tau \approx |\tau_0|/\gamma_0$. Estimates for ℓ_τ can be obtained from line tension measurements like those in [20], and range from 10^{-8} to 10^{-6} m. The ratio

$$\xi := \frac{\tau_0}{\gamma_0} \tag{2.3}$$

will be employed in the application shown in section 4. Finally, we can define a length ℓ_β as the typical size of a droplet for which

$$|\tau_0| \int_{\mathcal{C}} \mathrm{d}\ell \equiv \beta \int_{\mathcal{C}} \sigma^2 \mathrm{d}\ell$$

so that $\ell_\beta \approx \sqrt{\beta/|\tau_0|}$. We are unaware of any measure or estimate of ℓ_β. Although it might be reasonable to assume $\ell_\beta \ll \ell_\tau$, we will not make such a restriction in this paper. In any case, we do not need to apply our model up to the small lengths discussed here to appreciate the effects of line energies. As we will see in section 4, for instance, β could affect the stability of equilibrium configurations both quantitatively and qualitatively, even if it does not modify the equilibrium profile of $\mathcal{C}$ at all.

3. Equilibrium and stability. The first and the second variation of the free-energy functional $\mathcal{F}$ in (2.1) have been computed in [13] for $\beta = 0$. Here we simply arrive at the first variation $\delta\mathcal{F}^*$ of the reduced functional

$$\mathcal{F}^*[\mathcal{C}] := \int_{\mathcal{C}} \sigma^2 \mathrm{d}\ell. \tag{3.1}$$

The equilibrium equation obeyed by the droplet $\mathcal{B}$ is obtained by adding $\beta\delta\mathcal{F}^*$ to the Young equation (equation $(2.44)_2$ of [13]) specialized to the case where both the surface and the bare line tension are constant. It will be useful to write the functional $\mathcal{F}$ as

$$\mathcal{F} = \mathcal{F}_0 + \beta\mathcal{F}^*,$$

where $\mathcal{F}_0$ is the free-energy functional when $\beta = 0$.

Since $\mathcal{F}^*$ is concentrated along the contact line, it cannot affect the equilibrium shape of the free surface $\mathcal{S}^*$ which, in the absence of bulk contributions, is a surface with constant mean curvature.

To compute $\delta\mathcal{F}^*$ we perturb $\mathcal{C}$ by mapping points $p \in \mathcal{C}$ into points

$$p \mapsto p_\varepsilon := p + \varepsilon\boldsymbol{u} + \varepsilon^2\boldsymbol{v}, \tag{3.2}$$

where the regular fields $\boldsymbol{u}$ and $\boldsymbol{v}$ are defined on $\partial\mathcal{B}$. Since we do not repeat the computations for the complete functional $\mathcal{F}$, here we can deal with the restrictions of these fields along $\mathcal{C}$. In general, $\boldsymbol{u}$ and $\boldsymbol{v}$ are subject to the constraints [13]

$$\boldsymbol{u}\cdot\boldsymbol{\nu_*} = 0 \quad \text{and} \quad \boldsymbol{v}\cdot\boldsymbol{\nu_*} = -\frac{1}{2}\boldsymbol{u}\cdot(\nabla_{\mathrm{s}}\boldsymbol{\nu_*})\boldsymbol{u} \quad \text{on } \mathcal{S}_*, \tag{3.3}$$

where $\nabla_{\mathrm{s}}\boldsymbol{\nu_*} := (\nabla\boldsymbol{\nu_*})(\boldsymbol{I} - \boldsymbol{\nu_*}\otimes\boldsymbol{\nu_*})$ is the surface gradient of the outer unit normal $\boldsymbol{\nu_*}$ of $\mathcal{S}_*$. Equations (3.3) guarantee that the perturbed contact line glides on the substrate both at the first-order—$(3.3)_1$—and at the second-order—$(3.3)_2$—in the perturbation parameter ε. The field $\boldsymbol{v}$ does not enter into the equilibrium equations, but it plays a crucial role into the stability of the equilibrium configurations.

Let s be the arc-length of the contact line $\mathcal{C}$, and $\boldsymbol{t}_*$ its unit-tangent vector. We will frequently use the Darboux trihedron associated with $\mathcal{C}$: it is the set $\{\boldsymbol{t}_*, \boldsymbol{\nu}_*, \boldsymbol{\nu}_{\mathcal{S}_*}\}$ formed by three orthogonal unit vectors: $\boldsymbol{t}_*$, $\boldsymbol{\nu}_*$, and $\boldsymbol{\nu}_{\mathcal{S}_*} := \boldsymbol{t}_* \wedge \boldsymbol{\nu}_*$, the *conormal* unit vector of $\mathcal{C}$ on $\mathcal{S}_*$ (see Figure 1). When a point moves along $\mathcal{C}$, the associated Darboux trihedron obeys the following Darboux equations (see p. 241 of [26]):

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{t}_*}{\mathrm{d}s} = \kappa_g^*\boldsymbol{\nu}_{\mathcal{S}_*} + \kappa_n^*\boldsymbol{\nu}_*, \\[2ex] \dfrac{\mathrm{d}\boldsymbol{\nu}_{\mathcal{S}_*}}{\mathrm{d}s} = -\kappa_g^*\boldsymbol{t}_* - \tau_g^*\boldsymbol{\nu}_*, \\[2ex] \dfrac{\mathrm{d}\boldsymbol{\nu}_*}{\mathrm{d}s} = -\kappa_n^*\boldsymbol{t}_* + \tau_g^*\boldsymbol{\nu}_{\mathcal{S}_*}, \end{cases} \tag{3.4}$$

where

$$\kappa_n^* := \frac{\mathrm{d}\boldsymbol{t}_*}{\mathrm{d}s}\cdot\boldsymbol{\nu}_*, \quad \kappa_g^* := \frac{\mathrm{d}\boldsymbol{t}_*}{\mathrm{d}s}\cdot\boldsymbol{\nu}_{\mathcal{S}_*}, \quad \text{and} \quad \tau_g^* := \frac{\mathrm{d}\boldsymbol{\nu}_*}{\mathrm{d}s}\cdot\boldsymbol{\nu}_{\mathcal{S}_*} \tag{3.5}$$

are, respectively, the *normal curvature*, the *geodesic curvature*, and the *geodesic torsion* of $\mathcal{C}$, viewed as a curve on the substrate $\mathcal{S}_*$. Hereafter, to avoid clutter, we keep the star * only when we are referring to the unit normal $\boldsymbol{\nu}_*$ of $\mathcal{S}_*$, and when confusion might occur. No ambiguity should arise, since we always imagine $\mathcal{C}$ as a curve on $\mathcal{S}_*$.

By (3.2), we obtain

$$\frac{\mathrm{d}p_\varepsilon}{\mathrm{d}s} = \boldsymbol{t} + \varepsilon \boldsymbol{u}' + \varepsilon^2 \boldsymbol{v}'$$

and so

$$\frac{\mathrm{d}s}{\mathrm{d}s_\varepsilon} = \left|\frac{\mathrm{d}p_\varepsilon}{\mathrm{d}s}\right|^{-1} = [1 + 2\varepsilon \boldsymbol{u}\cdot\boldsymbol{t} + \varepsilon^2(\boldsymbol{u}'\cdot\boldsymbol{u}' + 2\boldsymbol{v}'\cdot\boldsymbol{t})]^{-1/2},$$

where a prime denotes differentiation with respect to s. Since

$$\boldsymbol{t}_\varepsilon = \frac{\mathrm{d}p_\varepsilon}{\mathrm{d}s_\varepsilon} = \frac{\mathrm{d}p_\varepsilon}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}s_\varepsilon},$$

it follows that

$$\boldsymbol{t}_\varepsilon = \boldsymbol{t} + \varepsilon[\boldsymbol{u}' - (\boldsymbol{u}'\cdot\boldsymbol{t})\boldsymbol{t}] + \varepsilon^2\left[\boldsymbol{v}' - \frac{1}{2}(\boldsymbol{u}'\cdot\boldsymbol{u}')\boldsymbol{t} - (\boldsymbol{v}'\cdot\boldsymbol{t})\boldsymbol{t} + \frac{3}{2}(\boldsymbol{u}'\cdot\boldsymbol{t})^2\boldsymbol{t} - (\boldsymbol{u}'\cdot\boldsymbol{t})\boldsymbol{u}'\right] + O(\varepsilon^3).$$

We introduce the vector fields

$$\boldsymbol{a} := \boldsymbol{u}' - (\boldsymbol{u}'\cdot\boldsymbol{t})\boldsymbol{t} \quad \text{and} \quad \boldsymbol{c} := \boldsymbol{v}' - (\boldsymbol{v}'\cdot\boldsymbol{t})\boldsymbol{t} \tag{3.6}$$

that satisfy $\boldsymbol{a}\cdot\boldsymbol{t} = \boldsymbol{c}\cdot\boldsymbol{t} = 0$. By setting $a^2 := \boldsymbol{a}\cdot\boldsymbol{a}$, we have $\boldsymbol{u}'\cdot\boldsymbol{u}' = a^2 + (\boldsymbol{u}'\cdot\boldsymbol{t})^2$ and so we can recast $\boldsymbol{t}_\varepsilon$ as

$$\boldsymbol{t}_\varepsilon = \boldsymbol{t} + \varepsilon\boldsymbol{a} + \varepsilon^2\left[\boldsymbol{c} - \frac{a^2}{2}\boldsymbol{t} - (\boldsymbol{u}'\cdot\boldsymbol{t})\boldsymbol{a}\right] + O(\varepsilon^3). \tag{3.7}$$

For a regular curve, the first Frenet–Serret equation states that

$$\frac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}s} = \sigma\boldsymbol{n}, \tag{3.8}$$

where $\boldsymbol{n}$ is the principal unit normal to the curve. Hence, on $\mathcal{C}_\varepsilon$ we have

$$\sigma_\varepsilon = \left[\frac{\mathrm{d}\boldsymbol{t}_\varepsilon}{\mathrm{d}s_\varepsilon}\cdot\frac{\mathrm{d}\boldsymbol{t}_\varepsilon}{\mathrm{d}s_\varepsilon}\right]^{1/2}$$

which, after rearrangements, yields

$$\sigma_\varepsilon^2\frac{\mathrm{d}s_\varepsilon}{\mathrm{d}s} = \frac{\mathrm{d}\boldsymbol{t}_\varepsilon}{\mathrm{d}s}\cdot\frac{\mathrm{d}\boldsymbol{t}_\varepsilon}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}s_\varepsilon}. \tag{3.9}$$

By use of (3.7) and (3.8) and after tedious but straightforward computations we obtain

$$\begin{aligned}\sigma_\varepsilon^2\frac{\mathrm{d}s_\varepsilon}{\mathrm{d}s} &= \sigma^2 + \varepsilon\Big[2\sigma\boldsymbol{a}'\cdot\boldsymbol{n} - \sigma^2(\boldsymbol{u}'\cdot\boldsymbol{t})] + \varepsilon^2[\boldsymbol{a}'\cdot\boldsymbol{a}' + 2\sigma\boldsymbol{n}\cdot\boldsymbol{c}' - \\ &-2\sigma(\boldsymbol{u}'\cdot\boldsymbol{t})'\boldsymbol{n}\cdot\boldsymbol{a} - 4\sigma(\boldsymbol{u}'\cdot\boldsymbol{t})\boldsymbol{n}\cdot\boldsymbol{a}' - \frac{3}{2}a^2\sigma^2 + \sigma^2(\boldsymbol{u}'\cdot\boldsymbol{t})^2 - \sigma^2\boldsymbol{v}'\cdot\boldsymbol{t}\Big].\end{aligned} \tag{3.10}$$

By definition, the first variation $\delta\mathcal{F}^*$ of $\mathcal{F}^*$ is given by

$$\delta\mathcal{F}^* := \left.\frac{\mathrm{d}\mathcal{F}^*[\mathcal{C}_\varepsilon]}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} = \int_{\mathcal{C}}[2\sigma\boldsymbol{a}'\cdot\boldsymbol{n} - \sigma^2(\boldsymbol{u}'\cdot\boldsymbol{t})]\mathrm{d}s = \int_{\mathcal{C}}[(\sigma^2\boldsymbol{t})'\cdot\boldsymbol{u} - 2(\sigma\boldsymbol{n})'\cdot\boldsymbol{a}]\mathrm{d}s, \tag{3.11}$$

where integration by parts has been used in the last passage. By recalling the definition of $\boldsymbol{a}$ in (3.6) and by performing several integrations by parts to get rid of the derivatives $\boldsymbol{u}'$, we obtain

$$\delta\mathcal{F}^* = \int_{\mathcal{C}}\{(\sigma^2\boldsymbol{t})' + 2(\sigma\boldsymbol{n})'' - 2[((\sigma\boldsymbol{n})'\cdot\boldsymbol{t})\boldsymbol{t}]'\}\cdot\boldsymbol{u}\mathrm{d}s.$$

Since the second Frenet–Serret equation reads

$$\frac{\mathrm{d}\boldsymbol{n}}{\mathrm{d}s} = -(\sigma\boldsymbol{t} + \widetilde{\tau}\boldsymbol{b}),$$

where $\widetilde{\tau}$ and $\boldsymbol{b} := \boldsymbol{t}\wedge\boldsymbol{n}$ are the torsion and the unit binormal vector of $\mathcal{C}$, we finally arrive at

$$\delta\mathcal{F}^* = \int_{\mathcal{C}}\boldsymbol{u}\cdot[3(\sigma^2\boldsymbol{t})' + 2(\sigma\boldsymbol{n})'']\mathrm{d}s.$$

The differential properties of $\mathcal{C}$ as a curve on $\mathcal{S}_*$ enter the scene when $(3.4)_1$ is compared with (3.8) so that $\delta\mathcal{F}^*$ reads as

$$\delta\mathcal{F}^* = \int_{\mathcal{C}}\boldsymbol{u}\cdot[6\sigma\sigma'\boldsymbol{t} + 3\sigma^3\boldsymbol{n} + 2(\kappa_g\boldsymbol{\nu}_{\mathcal{S}_*} + \kappa_n\boldsymbol{\nu}_{\boldsymbol{*}})'']\mathrm{d}s.$$

Since, by $(3.3)_1$,

$$\boldsymbol{u} = u_t\boldsymbol{t} + u_s\boldsymbol{\nu}_{\mathcal{S}_*} \tag{3.12}$$

along $\mathcal{C}$, by applying repeatedly the Darboux equations (3.4), by performing several integrations by parts, and by using the identity

$$\sigma^2 = \kappa_g^2 + \kappa_n^2, \tag{3.13}$$

we obtain

$$\begin{aligned}\delta\mathcal{F}^* = \int_{\mathcal{C}}\{[\kappa_g\sigma^2 + 2(\tau_g'\kappa_n + 2\tau_g\kappa_n' + \kappa_g'' - \kappa_g\tau_g^2)]u_s\\ + 2[\sigma\sigma' - \kappa_g\kappa_g' - \kappa_n\kappa_n']u_t\}\mathrm{d}s = \int_{\mathcal{C}}[\kappa_g\sigma^2 + 2(\tau_g'\kappa_n + 2\tau_g\kappa_n' + \kappa_g'' - \kappa_g\tau_g^2)]u_s\mathrm{d}s,\end{aligned}$$

where (3.13) was differentiated with respect to s to suppress the term multiplying u_t. The first variation $\delta\mathcal{F}^*$ is thus independent of the component u_t of $\boldsymbol{u}$ along the unit-tangent vector $\boldsymbol{t}$ of $\mathcal{C}$, as it should be, since u_t simply reparameterizes $\mathcal{C}$. One could assume a pragmatic attitude by setting $u_t \equiv 0$ from the very beginning. We prefer to keep this term since its disappearance from both the first and the second variation serves as a check of consistency for our computations.

If $\beta\delta\mathcal{F}^*$ is added to the first variation of $\mathcal{F}_0$ as given in (2.44) of [13], the following equilibrium equation should be obeyed along $\mathcal{C}$:

$$\gamma_0\cos\vartheta_c + \gamma_0 - w - \tau_0\kappa_g + \beta\kappa_g\sigma^2 + 2\beta(\tau_g'\kappa_n + 2\tau_g\kappa_n' + \kappa_g'' - \kappa_g\tau_g^2) = 0, \tag{3.14}$$

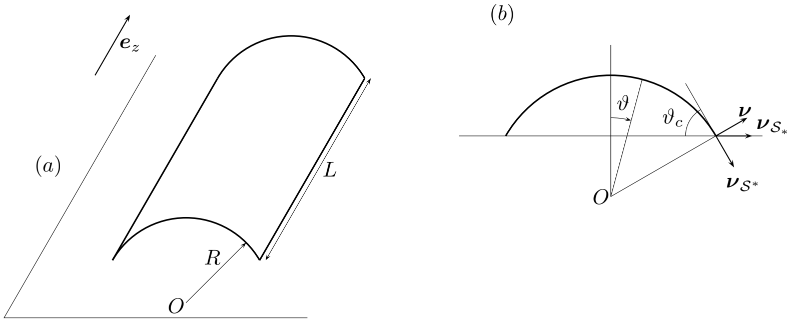


FIG. 2. (a) *Sketch of a liquid bridge, conceived as a straight circular cylinder of radius R, with symmetry axis along $\mathbf{e}_z$. The bridge is laid on a flat substrate. Here L denotes the typical length along which the cylinder is perturbed.* (b) *The cylindric polar coordinates z and ϑ used to parameterize the free surface of the bridge are shown together with the contact angle ϑ_c, which is constant along the contact line. The conormal unit vectors $\boldsymbol{\nu}_{\mathcal{S}^*}$ and $\boldsymbol{\nu}_{\mathcal{S}_*}$ of $\mathcal{C}$ as a curve on either the free or the adhering surface of the bridge have been drawn together with the unit normal vector $\boldsymbol{\nu}$ of the free surface along $\mathcal{C}$.*

where ϑ_c is the *contact angle*, that is, the angle between the conormal unit vectors $\boldsymbol{\nu}_{\mathcal{S}^*}$ and $\boldsymbol{\nu}_{\mathcal{S}_*}$ of $\mathcal{C}$ viewed as a curve on either $\mathcal{S}^*$ or $\mathcal{S}_*$, respectively (see Figure 1). At variance with (2.44) of [13], the subscript $_*$ has been dropped since no confusion can arise here. Although we will not study (3.14) in general, we note that setting $\beta \neq 0$ makes the curvature of the contact line appear at higher powers. This suggests that new branches of solutions might exist in this case.

The format just employed also gives the second variation of $\mathcal{F}^*$. Since computations are much more involved, however, we prefer to move the details into an appendix, while here we simply record the final result:

(3.15)
$$\begin{aligned}\delta^2\mathcal{F}^* \;&= \int_{\mathcal{C}}(u_s'')^2\mathrm{d}s + \int_{\mathcal{C}}(6\tau_g^2 - \kappa_g^2 - \tfrac{3}{2}\sigma^2)(u_s')^2\mathrm{d}s + \int_{\mathcal{C}}\Big\{\tau_g^4 + (\tau_g')^2 + \sigma^2(\kappa_g^2 - \tfrac{3}{2}\tau_g^2)\\ &\quad + (\kappa_n\tau_g)^2 + 2\kappa_n\kappa_g'\tau_g - 4\tau_g^2\kappa_g^2 + 4\kappa_g\tau_g\kappa_n' + [2\tau_g\kappa_g\kappa_n + 3\kappa_g\kappa_g']'\\ &\quad + (H-\kappa_n)(\frac{1}{2}\sigma^2\kappa_n + \kappa_n'' - 2\tau_g\kappa_g' - \kappa_g\tau_g' - \kappa_n\tau_g^2)\Big\}u_s^2\mathrm{d}s,\end{aligned}$$

where H is the total curvature of $\mathcal{S}_*$. By adding $\beta\delta^2\mathcal{F}^*$ to (3.16) of [13] we obtain the complete second variation of the functional $\mathcal{F}$. We remind the reader that u_{s*} in [13] coincides with u_s employed here.

4. Application. We apply the results of the previous sections to study the stability of a liquid bridge, conceived as a straight circular cylinder with radius R laid on a flat substrate (see Figure 2(a)). Since a cylinder is a surface with constant mean curvature, it represents an admissible equilibrium free surface. By (3.14), we see that a straight equilibrium contact line is unaffected by both the line tension and its curvature correction: the contact angle has a constant value ϑ_c along $\mathcal{C}$. We assume that the cylinder's axis lies along the $\boldsymbol{e}_z$ direction and we parameterize the free surface of the cylinder by using the angle $\vartheta \in [-\vartheta_c, \vartheta_c]$ and $z \in \mathbb{R}$ (see Figure 2(b)). Since

$\sigma = \kappa_n = \kappa_g = \tau_g = 0$, by (3.15) we have

$$\beta\delta^2\mathcal{F}^* = \beta\int_{\mathcal{C}}(u_s'')^2\mathrm{d}s,$$

which, when added to the second variation of $\mathcal{F}_0$ (see equation (3) of [16])

$$\delta^2\mathcal{F}_0[\boldsymbol{u}] = \gamma_0\int_{\mathcal{S}^*}\left\{|\nabla_{\mathrm{s}}u_\nu|^2 - \frac{1}{R^2}u_\nu^2\right\}\mathrm{d}a + \int_{\mathcal{C}}\left\{\tau_0(u_s')^2 - \frac{\gamma_0}{R}\cos\vartheta_c\sin\vartheta_c u_s^2\right\}\mathrm{d}s,$$

yields the second variation of the functional $\mathcal{F}$

(4.1)

$$\delta^2\mathcal{F}[\boldsymbol{u}] = \gamma_0\int_{\mathcal{S}^*}\left\{|\nabla_{\mathrm{s}}u_\nu|^2 - \frac{1}{R^2}u_\nu^2\right\}\mathrm{d}a + \int_{\mathcal{C}}\left\{\beta(u_s'')^2 + \tau_0(u_s')^2 - \frac{\gamma_0}{R}\cos\vartheta_c\sin\vartheta_c u_s^2\right\}\mathrm{d}s.$$

We warn the reader that in [16] the line tension was denoted by γ, and the surface tension by τ. In (4.1), $\nabla_{\mathrm{s}}u_\nu = (\boldsymbol{I} - \boldsymbol{\nu}\otimes\boldsymbol{\nu})\nabla u_\nu$ is the surface gradient of the scalar field u_ν, the component of $\boldsymbol{u}$ along the outer unit normal vector of the free surface $\mathcal{S}^*$. Along $\mathcal{C}$, u_ν is related to u_s, the projection of $\boldsymbol{u}$ along $\boldsymbol{\nu}_{\mathcal{S}_*}$, through the equation (see Figure 2(b))

$$u_\nu = \sin\vartheta_c u_s \tag{4.2}$$

to satisfy the gliding constraint $(3.3)_1$. Equation (4.2) can be obtained by expanding $\boldsymbol{u}$ along $\mathcal{C}$ as

$$\boldsymbol{u} = u_t\boldsymbol{t} + u_\nu\boldsymbol{\nu} + \overline{u}_s\boldsymbol{\nu}_{\mathcal{S}^*}, \tag{4.3}$$

which, together with (3.12), yields (4.2) after the scalar product with $\boldsymbol{\nu}$ has been formed and the equation $\boldsymbol{\nu}\cdot\boldsymbol{\nu}_{\mathcal{S}^*} = 0$ has been used too. With the aid of (4.2), $\delta^2\mathcal{F}$ becomes a quadratic functional of u_ν and so, either its minimum is zero, or it is unbounded from below. To deal with finite minima, we minimize $\delta^2\mathcal{F}$ on the set of functions obeying the constraint

$$\int_{\mathcal{S}^*}u_\nu^2\mathrm{d}a = 1. \tag{4.4}$$

If the minimum of $\delta^2\mathcal{F}$ on this set is positive, $\delta^2\mathcal{F}$ is positive definite, and so the equilibrium configuration is locally stable, whereas if the minimum of $\delta^2\mathcal{F}$ on the set (4.4) is negative, the equilibrium configuration is unstable [13]. Since we assume that the liquid bridge is made of incompressible fluid, u_ν should also obey the constraint (2.2). At first-order in ε (2.2) amounts to requiring [13]

$$\int_{\mathcal{S}^*}u_\nu\mathrm{d}a = 0\,. \tag{4.5}$$

The constraint (4.5), together with its second order implementation (equation $(2.29)_2$ of [13]) have been used in [13, 16] to obtain both the first and the second variation of $\mathcal{F}_0$. Until now we did not need them since we dealt only with $\mathcal{F}^*$ which is unaffected by this constraint, as it is concentrated on $\mathcal{C}$. However, to proceed we need to study $\delta^2\mathcal{F}$ and so we have to enforce incompressibility as well. Precisely, the first-order requirement (4.5) is needed since we have to compute only the first variation of $\delta^2\mathcal{F}$. Hence, we minimize the quadratic functional

$$\mathcal{G}[u_\nu] := \delta^2\mathcal{F}[u_\nu] - \frac{\mu}{2}\int_{\mathcal{S}}u_\nu^2\mathrm{d}a + \lambda\int_{\mathcal{S}^*}u_\nu\mathrm{d}a,$$

where $\mu/2$ and λ are Lagrange multipliers corresponding to the constraints (4.4) and (4.5). The scalar field u_ν is perturbed according to

$$u_\nu \mapsto u_{\nu\varepsilon} := u_\nu + \varepsilon\, h,$$

where h is a regular scalar field. Here we focus on the contribution arising from curvature correction. Much in the spirit of Rayleigh instability, we imagine that the cylinder has infinite length and we call L the length of $\mathcal{C}$ over which perturbations are effective. As a consequence, we require

$$u_\nu(\vartheta, 0) = u_\nu(\vartheta, L) = 0, \qquad \forall\vartheta \in [-\vartheta_c, \vartheta_c] \tag{4.6}$$

so that h also has to vanish at $z = 0, L$. By setting $\chi := 1/\sin\vartheta_c$, using (4.2), and integrating by parts twice, we obtain

$$\int_{\mathcal{C}} (u''_{s\varepsilon})^2 \mathrm{d}\ell = \int_0^L (u''_{s\varepsilon})^2 \mathrm{d}z = \chi^2 \int_0^L (u''_\nu)^2 \mathrm{d}z + 2\chi^2 \int_0^L h\, u_\nu^{(\mathrm{iv})} \mathrm{d}z + u''_\nu(L)\, h'(L) - u''_\nu(0)\, h'(0), \tag{4.7}$$

where a prime stands for differentiation along the arc-length z of $\mathcal{C}$, and use of (4.6) has been made. Since in this case $\mathcal{C}$ is an open curve we need to require

$$u''_\nu(\vartheta, 0) = u''_\nu(\vartheta, L) \qquad \forall\vartheta \in [-\vartheta_c, \vartheta_c]\,. \tag{4.8}$$

By adding (4.7) to the terms of the first variation of $\mathcal{G}$ that were computed in equations (7), (8) of [16], we conclude that finding the minimum of $\mathcal{G}$ on the set (4.4) amounts to finding the smallest eigenvalue μ of the following problem:

$$\triangle_{\mathrm{s}} u_\nu + \left(\mu + \frac{1}{R^2}\right) u_\nu + \lambda = 0 \quad \text{on } \mathcal{S}^*, \tag{4.9}$$

$$\sin^2\vartheta_c \nabla_{\mathrm{s}} u_\nu \cdot \boldsymbol{\nu}_{\mathcal{S}} + \frac{\beta}{\gamma_0} u_\nu^{(\mathrm{iv})} - \xi u''_\nu - \frac{1}{R}\sin\vartheta_c \cos\vartheta_c u_\nu = 0 \quad \text{along } \mathcal{C}, \tag{4.10}$$

where $\triangle_{\mathrm{s}}$ is the surface-Laplacian defined on $\mathcal{S}$ and ξ is defined according to (2.3). Hereafter we drop the subscript ν from u_ν. As proved in [13], the smallest value $\mu_{\min}$ of μ that solves the problem (4.9), (4.10) coincides with the minimum value of $\delta^2\mathcal{F}$ on the constraint (4.4) and so we conclude that an equilibrium configuration is locally stable or not according to whether $\mu_{\min}$ is positive or not.

To analyze (4.9), (4.10), we expand u as a sine series

$$u(\vartheta, z) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{2n\pi}{L} z\right) u_n(\vartheta)\,, \tag{4.11}$$

where $u_n(\vartheta)$ are unknown functions of ϑ. In this class, we can satisfy the boundary conditions (4.6) and (4.8) as well as the incompressibility constraint (4.5), so that we can set $\lambda = 0$ in (4.9). The reader might wonder whether the second variation just obtained, as well as that computed in [13, 16] for sessile droplets with *closed* contact line are valid here, where the contact line is open. A glance at the derivations of the second variation in the appendix and in [13, 16] shows that terms at the end-points of the contact line are *always* coupled with the curvature—normal or geodesic—or with the geodesic torsion of $\mathcal{C}$ which vanish identically along a straight contact line and so never contribute.

We split our discussion into two parts, according to whether $u_n(\vartheta)$ is symmetric with respect to the plane $\vartheta = 0$, or if it is skew-symmetric. We call *peristaltic* modes those in the former class, for which

$$\left.\frac{\partial u_n}{\partial \vartheta}\right|_{\vartheta=0} = 0 \quad \forall z \in [0, L] \tag{4.12}$$

holds and we call *varicose* the modes in the latter class, which in turn obey

$$u_n(0) = 0 \quad \forall z \in [0, L]\,. \tag{4.13}$$

4.1. Peristaltic modes. When a mode

$$u(\vartheta, z) = \sin\left(\frac{2n\pi}{L}z\right) u_n(\vartheta) \tag{4.14}$$

in the expansion (4.11) is inserted into (4.9) with $\lambda = 0$ and the multiplier μ is scaled to R^2, $u_n(\vartheta)$ has to satisfy

$$\frac{1}{R^2}\ddot{u}_n - \left(\frac{2n\pi}{L}\right)^2 u_n + \left(\frac{\mu+1}{R^2}\right) u_n = 0,$$

where we exploited the expression

$$\triangle_{\mathrm{s}} f = \frac{1}{R^2}\frac{\partial^2 f}{\partial \vartheta^2} + \frac{\partial^2 f}{\partial z^2}$$

of the surface-Laplacian acting on a scalar function $f = f(\vartheta, z)$ defined on $\mathcal{S}^*$ and where a superimposed dot denotes differentiation with respect to ϑ. The dimensionless ratio

$$\varrho_n := \left(\frac{2\pi n R}{L}\right)^2 \tag{4.15}$$

is a relative measure of the typical size R of a cross-section of the liquid bridge compared to the wavelength $L/2\pi n$ induced on $\mathcal{C}$ by the perturbation (4.14). For given R and L, the larger the ϱ_n, the more corrugated the contact line. By (4.12) and setting

$$\sigma_n := \mu + 1 - \varrho_n, \tag{4.16}$$

the peristaltic modes are given by

$$u_n(\vartheta) = \begin{cases} A\cos(\sqrt{\sigma_n}\vartheta) & \text{if } \sigma_n > 0, \\ A & \text{if } \sigma_n = 0, \\ A\cosh(\sqrt{-\sigma_n}\vartheta) & \text{if } \sigma_n < 0, \end{cases} \tag{4.17}$$

where A is an inessential constant that can be adjusted by imposing the constraint (4.4). If the mode (4.14) is inserted into (4.10), and we use (4.17), we conclude that $u(\vartheta, z)$ is an acceptable eigenfunction if

(4.18)
$$\frac{\beta}{\gamma_0}\left(\frac{2n\pi}{L}\right)^4 u_n(\vartheta_c) + \frac{1}{R}\dot{u}_n(\vartheta_c) + \xi\left(\frac{2n\pi}{L}\right)^2 \varrho_n u_n(\vartheta_c) - \frac{1}{R}u_n(\vartheta_c)\sin\vartheta_c\cos\vartheta_c = 0$$

holds, where we noted that

$$\nabla_{\mathrm{s}} u \cdot \boldsymbol{\nu}_{\mathcal{S}} = \frac{\partial u}{\partial \vartheta} = \dot{u}(\vartheta)\,.$$

By recalling (2.3) we introduce the dimensionless parameters

$$\omega := \frac{\xi}{R} = \frac{\tau_0}{\gamma_0 R} \qquad \text{and} \qquad \eta := \frac{\beta}{\gamma_0 R^3}$$

that compare the length scales $|\tau_0|/\gamma_0$ and $\sqrt[3]{\beta/\gamma_0}$ hidden in the model with R, a characteristic length of the liquid bridge. By setting

$$x_n := \begin{cases} \sqrt{\sigma_n}\vartheta_c & \text{if } \sigma_n > 0, \\ \sqrt{-\sigma_n}\vartheta_c & \text{if } \sigma_n < 0, \end{cases} \tag{4.19}$$

we can recast (4.18) as

$$\eta\varrho_n^2 + \omega\varrho_n = \sin\vartheta_c \left[\cos\vartheta_c + \frac{\sin\vartheta_c}{\vartheta_c} x_n \tan x_n\right] \quad \text{if } \sigma_n > 0, \tag{4.20}$$

$$\eta\varrho_n^2 + \omega\varrho_n = \sin\vartheta_c \left[\cos\vartheta_c - \frac{\sin\vartheta_c}{\vartheta_c} x_n \tanh x_n\right] \quad \text{if } \sigma_n < 0, \tag{4.21}$$

and

$$\eta\varrho_n^2 + \omega\varrho_n = \sin\vartheta_c \cos\vartheta_c \quad \text{if } \sigma_n = 0\,. \tag{4.22}$$

Following [16], modes satisfying (4.20), (4.21), and (4.22) are called *circular*, *hyperbolic*, and *linear* modes, respectively. Compared to the analysis performed in [16], the left-hand side of (4.20)–(4.22) is a second- instead of a first-degree polynomial in ϱ_n: in this sense, the curvature correction acts as a singular perturbation term. As we remarked before, stable modes correspond to positive values of μ, whereas unstable modes correspond to $\mu < 0$. To ascertain the stability of a particular mode, it is then crucial to localize the *marginal modes*, corresponding to $\mu = 0$. By the definitions (4.16) and (4.19) we can write

$$\mu = \begin{cases} \varrho_n - 1 + \left(\frac{x_n}{\vartheta_c}\right)^2 & \text{if } \sigma_n > 0, \\ \varrho_n - 1 - \left(\frac{x_n}{\vartheta_c}\right)^2 & \text{if } \sigma_n < 0, \\ \varrho_n - 1 & \text{if } \sigma_n = 0\,. \end{cases} \tag{4.23}$$

From $(4.23)_1$ we conclude that points in the quadrant $\mathcal{Q} := \{(x_n, \varrho_n) \mid x_n \geq 0,\, \varrho_n \geq 0\}$ of the (x_n, ϱ_n)-plane that lie below the parabola

$$\varrho_n = 1 - \left(\frac{x_n}{\vartheta_c}\right)^2 \tag{4.24}$$

are unstable against circular modes, whereas points in $\mathcal{Q}$ above this parabola are stable against circular modes. Similarly, it follows from $(4.23)_2$ that points of $\mathcal{Q}$ below the parabola

$$\varrho_n = 1 + \left(\frac{x_n}{\vartheta_c}\right)^2 \tag{4.25}$$

are unstable against hyperbolic modes, whereas points above it are stable against hyperbolic modes. Finally, points in $\mathcal{Q}$ that lie below the straight line

$$\varrho_n = 1 \tag{4.26}$$

are unstable against linear modes, while points above this line are stable. We now replace ϱ_n in (4.20)–(4.22) with the appropriate expressions found in (4.24)–(4.26), and we divide (4.20)–(4.22) by ω and define the functions

$$g_c(x_n) := \phi\left\{\eta\left[1-\left(\frac{x_n}{\vartheta_c}\right)^2\right]^2 - \sin\vartheta_c[\cos\vartheta_c + \frac{\sin\vartheta_c}{\vartheta_c}x_n\tan x_n]\right\},$$

$$g_h(x_n) := \phi\left\{\eta\left[1+\left(\frac{x_n}{\vartheta_c}\right)^2\right]^2 - \sin\vartheta_c[\cos\vartheta_c - \frac{\sin\vartheta_c}{\vartheta_c}x_n\tanh x_n]\right\},$$

and

$$g_l(x_n) := \phi[\eta - \sin\vartheta_c\cos\vartheta_c],$$

where, for simplicity, we set $\phi := 1/|\omega|$. Increasing values of ϕ correspond to line tensions with decreasing magnitude. By (4.23), the marginal modes are the smaller pairs (x_n, ϱ_n) in $\mathcal{Q}$ that obey the equation

$$\begin{cases} 1-\left(\frac{x_n}{\vartheta_c}\right)^2 = g_c(x_n) & \text{if } \sigma_n > 0, \\ 1+\left(\frac{x_n}{\vartheta_c}\right)^2 = g_h(x_n) & \text{if } \sigma_n < 0, \\ 1 = g_l(x_n) & \text{if } \sigma_n = 0. \end{cases} \tag{4.27}$$

The pairs (ϕ, ϱ_n) that solve (4.27) and yield the most restrictive stability condition lie on the *marginal* curve which divides the (ϕ, ϱ_n)-plane into a stable and an unstable set. Figure 3 shows the marginal curves for $\vartheta_c = 65^\circ$ and for several values of η: no qualitative differences occur if other values of $\vartheta_c < \pi/2$ are chosen. To follow the discussion the reader is also urged to look at Figure 4, where the semilogarithmic plot of the marginal curves of Figure 3 is shown. The numerical solution of (4.27) (and of (4.28) below) has been performed in a Matlab environment by resorting to a simple bisection algorithm. An educated guess based on an a priori analytical study of the equation has been used to select the intervals in which the solutions are first sought. We stress that, by definition of ϕ and since we consider only negative line tensions, moving along the ϕ-axis from left to right amounts to spanning the interval $(-\infty, 0)$ for the line tension.

Hyperbolic modes are most effective and, depending on the value of η, we can single out three stability diagrams, according to the profile of the marginal curve. If $\eta = 0$ the marginal curve has a turning point and the stability diagram coincides with that obtained in [16]. A straight line $\phi = \phi_0 =$ const. either intersects the marginal curve twice or it does not intersect it at all, according to the value of ϕ_0. When two intersections $(\phi_0, \varrho_n^{(1)})$ and $(\phi_0, \varrho_n^{(2)})$ exist $(\varrho_n^{(1)} < \varrho_n^{(2)})$, Rayleigh instability makes liquid bridges unstable when $\varrho_n < \varrho_n^{(1)}$. When $\varrho_n \in (\varrho_n^{(1)}, \varrho_n^{(2)})$ liquid bridges are locally stable and they become unstable again when $\varrho_n > \varrho_n^{(2)}$. Since ϱ_n is proportional to n, modes with $n = 1$ are more likely to induce Rayleigh instability:

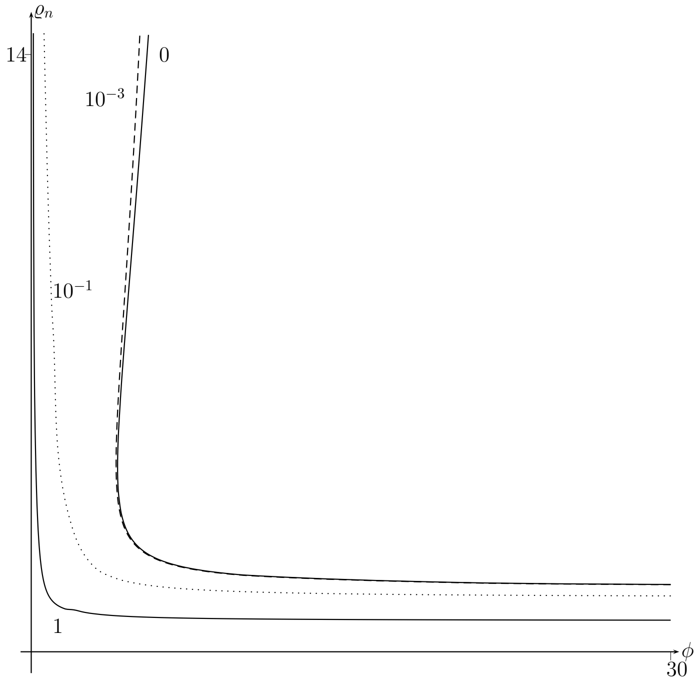


FIG. 3. *Stability diagram for peristaltic modes when* $\vartheta_c = 65°$*. Each curve is labelled by the corresponding value of* η*:* $\eta = 0$*,* $\eta = 10^{-3}$*,* $\eta = 10^{-1}$*, and* $\eta = 1$*). Only hyperbolic modes are effective. For a given value of* η*, the region bounded by the coordinate axes and the marginal curve is unstable, while the remaining portion of the* (ϕ, ϱ_n)*-plane is stable. On increasing* η*, Rayleigh instability persists while the instability induced by negative line tension for large values of* n *is reduced, since the marginal curve diverges only in the limit as* $\phi \to 0$*, that is, when the magnitude of line tension is exceedingly high (see also Figure* 4 *for further details).*

in fact, we leave the unstable set $\varrho_n < \varrho_n^{(1)}$ earlier and earlier on increasing n at fixed L and R. The effects of line tension are more related to the unstable set $\varrho_n > \varrho_n^{(2)}$. In this case, modes with large n are most likely to cause instability. However [16], n cannot be increased arbitrarily in a coherent theory since the typical length scale L/n associated with the corrugations induced on the contact line by the perturbation falls *below* the smallest scale that can be reached within a continuum approach. Hence, only a finite number of values of n can be considered and it is clear from Figure 3 that the smaller the line tension, the more values of n will fall within the region of local stability. When $\phi = \phi_0$ does not intersect the marginal curve, and so the line tension has a large magnitude, *no* stable modes survive, and no stable equilibrium liquid bridge exists.

If $\eta \in (0, \eta_c(\vartheta_c)]$ (curves labelled by $\eta = 10^{-3}$ in Figures 3 and 4) the marginal curve has two turning points. A line $\phi = \phi_0$ crosses the marginal curve three times if $\phi_0 \in [\phi_0^m, \phi_0^M]$ and only once elsewhere. In this latter case only Rayleigh instability occurs: it is slightly reduced by a curvature correction when $\phi_0 > \phi_0^M$ (i.e., when line tension is small), but it becomes more and more restrictive if $\phi_0 < \phi_0^m$ (i.e., when line tension is large), since the marginal curve diverges along the ϱ_n axis. When $\phi_0 \in [\phi_0^m, \phi_0^M]$ the points $(\phi_0, \varrho_n^{(1)})$, $(\phi_0, \varrho_n^{(2)})$, and $(\phi_0, \varrho_n^{(3)})$ $(\varrho_n^{(1)} < \varrho_n^{(2)} < \varrho_n^{(3)})$ on the marginal curve impose the following scenario: a liquid bridge is unstable when either $\varrho_n < \varrho_n^{(1)}$ (Rayleigh instability) or $\varrho \in (\varrho_n^{(2)}, \varrho_n^{(3)})$ and stable when either

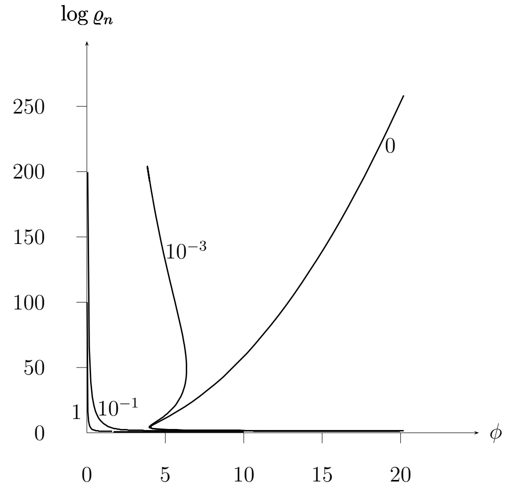


FIG. 4. *Semilogarithmic plot of the stability diagram for peristaltic modes shown in Figure* 3. *The four curves are labelled by the corresponding values of* η. *The region of Rayleigh instability can be perceived only for* $\eta = 10^{-1}$ *and* $\eta = 1$. *Three regimes exist, depending on the value of* η. *When* $\eta = 0$ *a line* $\phi = \phi_0$ *either crosses the marginal curve twice or it does not cross it at all. This mirrors the destabilizing role of negative line tension when there are no curvature corrections. The second regime covers the case* $\eta \in (0, \eta_c]$*: here,* $\eta = 10^{-3}$. *Then, when* ϕ *is either very small or large,* $\phi = \phi_0$ *crosses the marginal curve only once, and only Rayleigh instability occurs. There is an intermediate set of values of* ϕ_0 *for which three intersections exist between* $\phi = \phi_0$ *and the marginal curve. Finally, when* $\eta > \eta_c$*—here* $\eta = 10^{-1}$ *and* $\eta = 1$*—there is always one intersection between a line* $\phi = \phi_0$ *and the marginal curve: only Rayleigh instability occurs.*

$\varrho \in (\varrho_n^{(1)}, \varrho_n^{(2)})$ or $\varrho > \varrho_n^{(3)}$. In particular, the local stability when $\varrho > \varrho_n^{(3)}$ mirrors the stabilizing role of even a tiny curvature correction. The undulating behavior for $\varrho_n \in (\varrho_n^{(1)}, \varrho_n^{(3)})$ disappears when η attains a critical value $\eta_c(\vartheta_c)$ at which $\phi_0^m = \phi_0^M$. For larger values of η (curves labelled by $\eta = 10^{-1}$ and $\eta = 1$ in Figures 3 and 4) the marginal curve has a monotonic profile and it is crossed by a line $\phi = \phi_0$ at a unique point $(\phi_0, \varrho_n^{(1)})$: only liquid bridges such that $\varrho_n < \varrho_n^{(1)}$ are unstable: a Rayleigh instability occurs which becomes stronger and stronger when the magnitude of line tension increases (i.e., when $\phi \to 0^+$). Figure 5 shows the graph of $\eta_c(\vartheta_c)$ against the contact angle ϑ_c. We conclude that small and large values of ϑ_c require lower values of η to wash out the instability at large n typical of a negative line tension. To prove that the marginal curve cannot diverge in the limit $\phi \to \infty$ and when η assumes any *fixed*, nonvanishing value, we simply look at (4.27) *before* division by ω is performed. By applying the method of dominant balance [27], we conclude that $\varrho_n \to \infty$ and $\phi \to \infty$ would yield

$$\eta \varrho_n^2 = -\frac{\sin^2 \vartheta_c}{\vartheta_c} x_n \tanh x_n,$$

which is clearly inconsistent because of the different sign of the two sides. Similarly, we exclude that ϱ_n could diverge at a finite value of ϕ. Hence, given a fixed value of $\eta > 0$ ϱ_n could diverge only in the limit where $\phi \to 0$, that is, if the negative line tension has a large magnitude. This argument corroborates the outcomes of the numerical analysis of (4.27).

Figure 6 shows the stability diagram of a liquid bridge against peristaltic modes, when the contact angle is larger than $\pi/2$: precisely, here $\vartheta_c = 125°$. Both circular

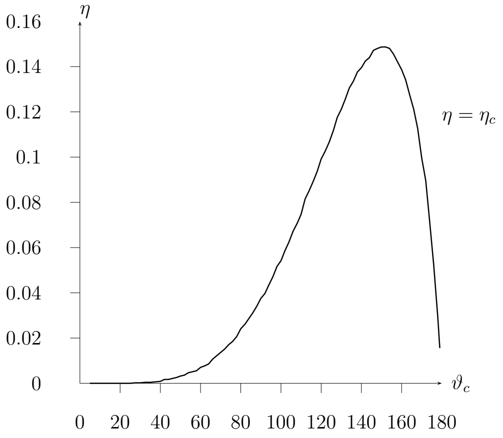


FIG. 5. *The critical value η_c of η is plotted against the contact angle ϑ_c for peristaltic modes. When η exceeds η_c, the marginal curve is a monotonic function of ϕ.*

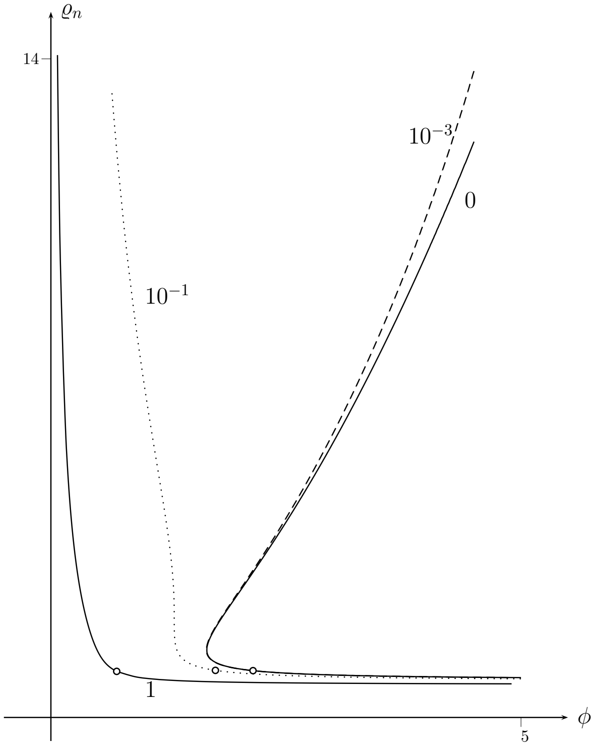


FIG. 6. *Stability diagram against peristaltic modes when $\vartheta_c = 125°$. The marginal curves plotted here are labelled with the corresponding values of η. Both circular and hyperbolic modes are effective in this case. The portion of a given marginal curve that lies above the circle consists of hyperbolic modes, while the portion below the circle consists of circular modes. Linear modes never affect the stability diagram. Apart from the presence of two families of modes, there is no qualitative difference with respect to the case $\vartheta_c < \pi/2$.*

and hyperbolic modes are effective in this case but, apart from this, there is no substantial difference from the case where $\vartheta_c < \pi/2$. Similarly, the semilogarithmic

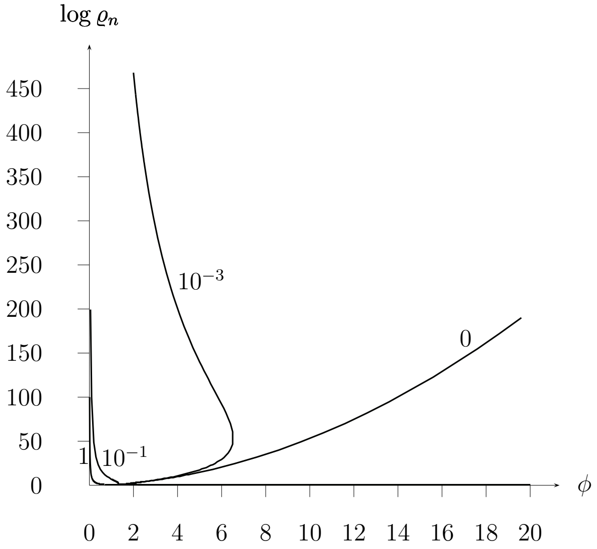


FIG. 7. *Semilogarithmic plot of the stability diagram for peristaltic modes shown in Figure* 6. *Each marginal curve is labelled by the corresponding value of* η.

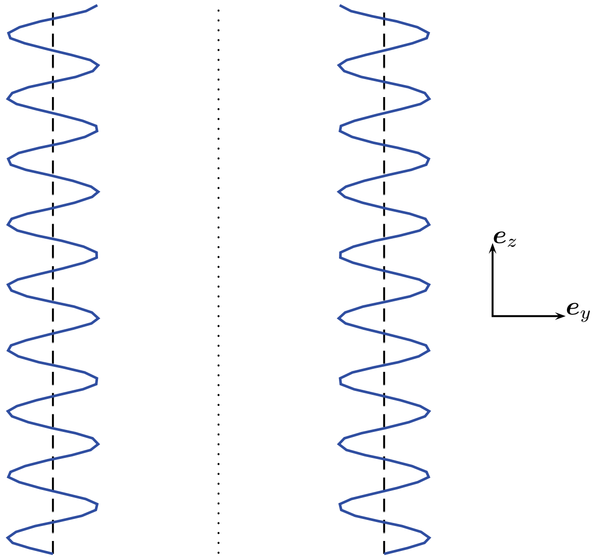


FIG. 8. *The unperturbed (dashed lines) and the perturbed (solid lines) contact lines of a marginally stable peristaltic mode are plotted here when* $\vartheta_c = 65°$, $R = 1$, *and* $L = 10$. *The two sinusoidal profiles are symmetric with respect to the axis of the bridge (dotted line).*

plot in Figure 7 does not have new features as compared with that in Figure 4.

Finally, Figures 8 and 9 show, respectively, the perturbed contact lines and the perturbed cross-section of a liquid bridge induced by a marginal mode. We chose an equilibrium liquid bridge with $\vartheta_c = 65°$, $R = 1$, and $L = 10$. The value of ϱ_n corresponding to the marginal mode is $\varrho_n = 30$ and the amplitude of the marginal mode has been magnified to appreciate its structure.

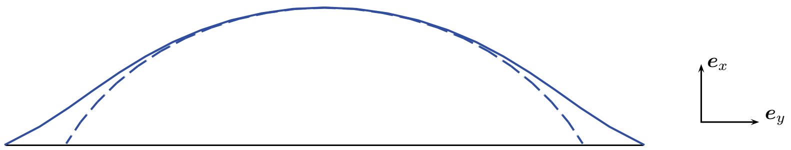


FIG. 9. *Cross-section in a plane $z = z_0$ of an unperturbed liquid bridge (dashed line) and of a marginally stable peristaltic mode (solid line). The geometric parameters are $\vartheta_c = 65°$, $R = 1$, and $L = 10$. For this value of z_0, the perturbed cross-section has larger area than the unperturbed one but, to obey the incompressibility constraint, the opposite is true for other values of z_0.*

To obtain the perturbed profile of the contact lines in Figure 8 we noted that the straight contact lines of the unperturbed bridge satisfy $y = \pm R\sin\vartheta_c$ and so, by (3.2), (4.2), and (4.14) we can write on them

$$p_\varepsilon - O = \pm\sin\vartheta_c \boldsymbol{e}_z \pm \varepsilon\frac{1}{\sin\vartheta_c}u_n(\pm\vartheta_c, z)\boldsymbol{e}_y\,.$$

Since $\nabla_s\boldsymbol{\nu}_* = \boldsymbol{0}$, by $(3.3)_2$ we have $\boldsymbol{v}\cdot\boldsymbol{\nu}_* = 0$ and so $\boldsymbol{v}$ in (3.2) is not needed here. To obtain the cross-section at $z = z_0$ of a perturbed profile we first note that the outer unit normal $\boldsymbol{\nu}$ to the unperturbed free surface is (see Figure 2(b))

$$\boldsymbol{\nu}(\vartheta) = \cos\vartheta\boldsymbol{e}_x + \sin\vartheta\boldsymbol{e}_y\,.$$

We then introduce a vector, still called $\boldsymbol{\nu}_{\mathcal{S}^*}$, which is tangent to the unperturbed cylinder and that coincides with the conormal unit vector $\boldsymbol{\nu}_{\mathcal{S}^*}$ at the contact line $y = R\sin\vartheta_c$, that is,

$$\boldsymbol{\nu}_{\mathcal{S}^*}(\vartheta) = \sin\vartheta\boldsymbol{e}_x - \cos\vartheta\boldsymbol{e}_y\,.$$

By combining (4.2) and (4.3) and recalling that $\boldsymbol{\nu}_{\mathcal{S}^*}\cdot\boldsymbol{\nu}_{\mathcal{S}_*} = \cos\vartheta_c$ along $\mathcal{C}$, we conclude that

$$\overline{u}_s(\vartheta_c, z) = u_s(\vartheta_c, z)\cos\vartheta_c = u_\nu(\vartheta_c, z)\cot\vartheta_c\,.$$

The value of $\overline{u}_s$ in the bulk is irrelevant since it simply amounts to a different parametrization of the cross-section. We are then free to select

$$\overline{u}_s(\vartheta, z) = \frac{\vartheta}{\vartheta_c}\overline{u}_s(\vartheta_c, z) \quad \vartheta\in[0,\vartheta_c]$$

that satisfies the only constraint it has to obey, namely to have a prescribed value along the contact line. Hence, the perturbed marginal profile is given by

$$p_\varepsilon - O = R\boldsymbol{\nu} + \varepsilon u(\vartheta, z_0)\boldsymbol{\nu} + \varepsilon\overline{u}_s(\vartheta, z)\boldsymbol{\nu}_{\mathcal{S}^*}(\vartheta) \quad \vartheta\in[0,\vartheta_c],$$

where $u(\vartheta, z)$ is given by (4.14). The profile of the perturbed cross-section for $\vartheta\in[-\vartheta_c, 0]$ is obtained by symmetry about the $\boldsymbol{e}_x$-axis.

4.2. Varicose modes. In this class, $u_n(0)\equiv 0$ and so, by retracing the same steps as before, we obtain

$$u_n = \begin{cases} A\sin(\sqrt{\sigma_n}\vartheta) & \text{if } \sigma_n > 0,\\ A\sinh(\sqrt{-\sigma_n}\vartheta) & \text{if } \sigma_n < 0,\end{cases}$$

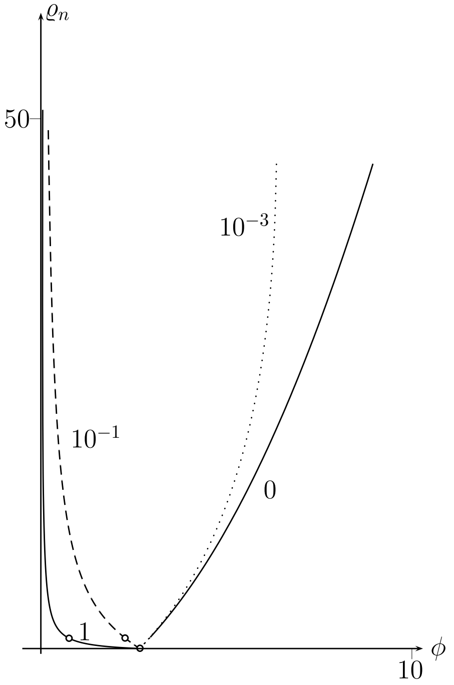


FIG. 10. *Stability diagrams for varicose—both circular and hyperbolic—modes. Here, $\vartheta_c = 65^\circ$ and the marginal curves are labelled by the corresponding values of η. The portion of a given marginal curve that lies above the circle consists of hyperbolic modes, while the portion below the circle consists of circular modes. The marginal curves coalesce along the ϕ axis, since $\varrho_n = 0$ always solves* (4.28)$_1$. *This solution does not cause instability since only positive values of ϱ_n are meaningful. For any given value of η the region bounded by the marginal curve and the ϱ_n-axis is unstable against varicose modes.*

for circular and hyperbolic modes, respectively, while linear modes are absent. From this point, the analysis of section 4.1 can be repeated *verbatim*. After introducing the functions

$$k_c(x_n) := \phi \left\{ \eta \left[1 - \left(\frac{x_n}{\vartheta_c} \right)^2 \right]^2 - \sin\vartheta_c [\cos\vartheta_c - \frac{\sin\vartheta_c}{\vartheta_c} x_n \cot x_n] \right\} ,$$

$$k_h(x_n) := \phi \left\{ \eta \left[1 + \left(\frac{x_n}{\vartheta_c} \right)^2 \right]^2 - \sin\vartheta_c [\cos\vartheta_c - \frac{\sin\vartheta_c}{\vartheta_c} x_n \coth x_n] \right\} ,$$

marginal modes are obtained by determining the smallest pairs in $\mathcal{Q}$ that obey

$$\left\{ \begin{array}{ll} 1 - \left(\frac{x_n}{\vartheta_c}\right)^2 = k_c(x_n) & \text{if } \sigma_n > 0 \\ 1 + \left(\frac{x_n}{\vartheta_c}\right)^2 = k_h(x_n) & \text{if } \sigma_n < 0 \, . \end{array} \right. \tag{4.28}$$

Figure 10 shows the stability diagram of a liquid bridge against varicose modes, when $\vartheta_c = 65^\circ$. The branch of the marginal curve corresponding to Rayleigh instability disappears. When $\phi \to \infty$, varicose modes are stable, as it should be, since they do

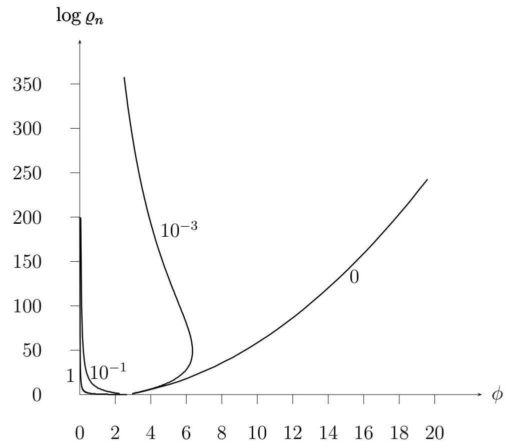


FIG. 11. *Semilogarithmic plot of the stability diagrams for varicose modes shown in Figure* 10.

not affect Rayleigh instability in the absence of line tension. Let us first consider the case $\eta = 0$. When the magnitude of negative line tension is progressively decreased, instability occurs for large values of ϱ_n. To grasp the behavior of the marginal curve in the limit where $\phi \to \infty$, we still employ the method of dominant balance. Since large values of ϱ_n also imply large values of x_n by (4.25), we can look for solutions to $(4.28)_2$ in the form $x_n = b\phi^\alpha$, where b and α are two positive numbers to be determined. When we replace this ansatz into $(4.28)_2$ and discard negligible terms, we arrive at

$$\frac{b}{\vartheta_c}\phi^\alpha \left[\frac{b}{\vartheta_c}\phi^\alpha - \phi \sin^2 \vartheta_c\right] = 0$$

whence $\alpha = 1$ and $b = \vartheta_c \sin^2 \vartheta_c$ follow. Figure 10 points out a difference between peristaltic and varicose modes since for these latter the marginal curves emanate from a precise point $\phi(\vartheta_c)$ of the ϕ axis: $\phi(65^\circ) = 2.66$. As for peristaltic modes, at any fixed, nonvanishing value for η, the marginal curve diverges along the ϱ_n axis (see the semilogarithmic plot shown in Figure 11), confirming the stabilizing role of curvature corrections. The same regimes discussed for peristaltic modes exist here, apart from the absence of Rayleigh instability when $\phi > \phi(\vartheta_c)$.

Figure 12 shows the stability diagram when $\vartheta_c = 125^\circ$. As already discussed for peristaltic modes, there are no essential differences with respect to the case $\vartheta_c \leq \pi/2$. Similar remarks hold for the semilogarithmic counterpart shown in Figure 13. Finally, Figure 14 shows the critical value of $\eta_c(\vartheta_c)$ at which the marginal curve follows a monotonic profile: it has the same qualitative behavior as that computed for peristaltic modes.

5. Conclusions. We determined the effects of a curvature correction to line tension on both the equilibrium and the stability of sessile droplets through a general variational analysis. While the effects on the equilibrium could even be absent, those on stability are relevant in any case. As a first consequence, we proved for liquid bridges that the curvature correction makes wildly oscillating perturbations unrewarding, and so the systematic instability against all modes with short wavelength induced by negative line tensions is removed, regardless of the magnitude of the correction. This magnitude, however, plays a crucial role in determining whether

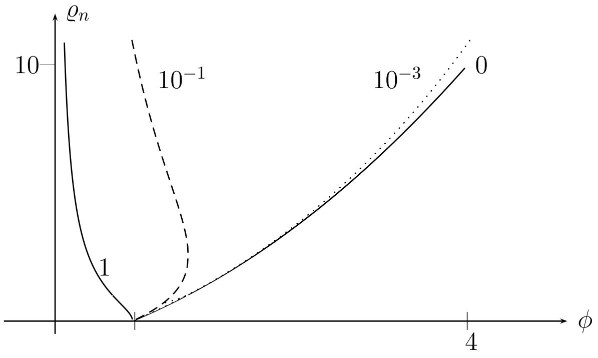


FIG. 12. *Stability diagram against varicose modes when $\vartheta_c = 125^\circ$. The marginal curves plotted here are labelled by the corresponding values of η. Both circular and hyperbolic modes are effective, but the transition between them is too close to the ϕ axis to be shown here. Linear modes never affect the stability diagram. Also in this case, the diagram is similar to that for the case $\vartheta_c < \pi/2$.*

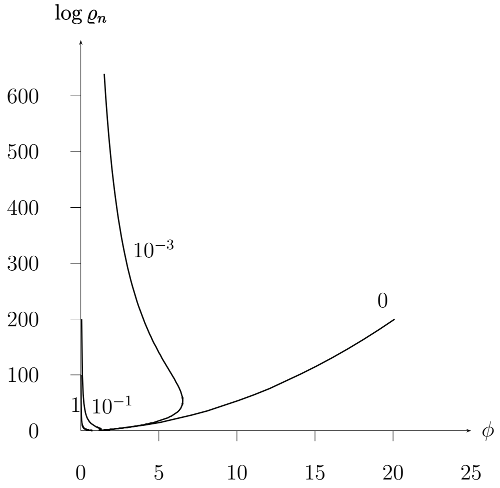


FIG. 13. *Semilogarithmic plot of the stability diagrams for varicose modes shown in Figure* 12. *The marginal curves are labelled by the corresponding values of η.*

only Rayleigh instability occurs or not. As a general result, Rayleigh instability is the only destabilizing mechanism whenever the curvature correction is large enough. The analysis employed here for liquid bridges could serve to explore the stabilizing effects of curvature corrections on droplets with a closed geometry like spherical caps. We expect that the stabilizing mechanism discussed here will work also in that context, although the computations will be more cumbersome.

Appendix. Second variation of $\mathcal{F}^*$. We show in detail how to arrive at the expression (3.15) for the second variation $\delta^2\mathcal{F}^*$ of $\mathcal{F}^*$, obtained by integrating along $\mathcal{C}$ the terms in (3.10) that are quadratic in ε. We start with

$$\mathcal{I}_1 := \int_{\mathcal{C}} [2\sigma \boldsymbol{c}' \cdot \boldsymbol{n} - \sigma' \boldsymbol{v} \cdot \boldsymbol{t}] \mathrm{d}s,$$

which contains contributions related to the field $\boldsymbol{v}$ defined in (3.2). The integral $\mathcal{I}_1$ has the same structure as the first variation of $\mathcal{F}^*$ given in (3.11), with $\boldsymbol{u}$ and $\boldsymbol{a}$ replaced

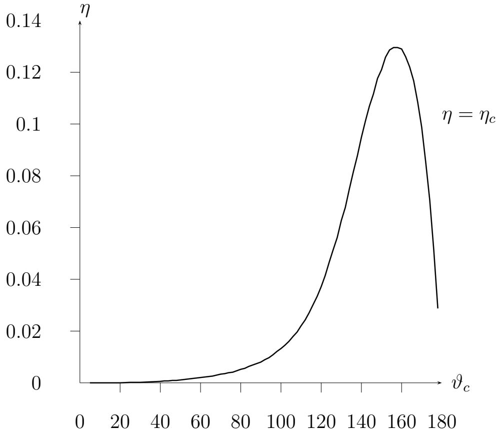


FIG. 14. *The critical value η_c of η is plotted against the contact angle ϑ_c for varicose modes. When η exceeds η_c, the marginal curve is a monotonic function of ϕ.*

by $\boldsymbol{v}$ and $\boldsymbol{c}$. The crucial difference in this formal change is that, at variance with $\boldsymbol{u}$, the field $\boldsymbol{v}$ has also a nontrivial component along the unit normal vector $\boldsymbol{\nu}_*$ of $\mathcal{S}_*$. By retracing the same steps as in section 3, we can check that the component $\boldsymbol{v}\cdot\boldsymbol{t}$ does not contribute and the component along $\boldsymbol{\nu}_{\mathcal{S}_*}$ vanishes by virtue of the equilibrium equation (3.14). Hence, we are left with the component $\boldsymbol{v}\cdot\boldsymbol{\nu}_*$ which, by use of $(3.3)_2$, can be recast as

$$\mathcal{I}_1 = -\int_{\mathcal{C}} \frac{1}{2}\boldsymbol{u}\cdot(\nabla_s\boldsymbol{\nu}_*)\boldsymbol{u}[\sigma^2\kappa_n + 2\kappa_n'' - 4\tau_g\kappa_g' - 2\kappa_g\tau_g' - 2\kappa_n\tau_g^2]\mathrm{d}s\,,$$

where perusal of Darboux equations (3.4) has been made. By recalling that [25]

$$\nabla_s\boldsymbol{\nu}_* = -\kappa_n\boldsymbol{t}\otimes\boldsymbol{t} - \kappa_{n\perp}\boldsymbol{\nu}_{\mathcal{S}}\otimes\boldsymbol{\nu}_{\mathcal{S}} + \tau_g(\boldsymbol{\nu}_{\mathcal{S}}\otimes\boldsymbol{t} + \boldsymbol{t}\otimes\boldsymbol{\nu}_{\mathcal{S}})\,, \tag{A.1}$$

where $\kappa_{n\perp} := H - \kappa_n$ is expressed in terms of the total curvature H of $\mathcal{S}_*$, we finally arrive at

$$\mathcal{I}_1 = \int_{\mathcal{C}} \frac{1}{2}[\kappa_n u_t^2 - 2\tau_g u_t u_s + (H-\kappa_n)u_s^2]\{\sigma^2\kappa_n + 2\kappa_n'' - 4\tau_g\kappa_g' - 2\kappa_g\tau_g' - 2\kappa_n\tau_g^2\}\mathrm{d}s. \tag{A.2}$$

It is also expedient to expand $\boldsymbol{u}'$ and $\boldsymbol{a}$ along the Darboux trihedron of $\mathcal{C}$, by resorting to (3.4), $(3.6)_1$, and (3.13):

$$\boldsymbol{u}' = (u_t' - \kappa_g u_s)\boldsymbol{t} + (u_s' + \kappa_g u_t)\boldsymbol{\nu}_{\mathcal{S}} + (\kappa_n u_t - \tau_g u_s)\boldsymbol{\nu} \tag{A.3}$$

and

$$\boldsymbol{a} = (u_s' + \kappa_g u_t)\boldsymbol{\nu}_{\mathcal{S}} + (\kappa_n u_t - \tau_g u_s)\boldsymbol{\nu}, \tag{A.4}$$

from which

$$a^2 = u_t^2\sigma^2 + u_s'^2 + \tau_g^2 u_s^2 + 2\kappa_g u_t u_s' - 2\tau_g\kappa_n u_t u_s$$

easily follows. By differentiating $\boldsymbol{a}$ with respect to s we also obtain, by (3.4),

$$\begin{aligned}\boldsymbol{a}' &= [\kappa_n\tau_g u_s - \kappa_g u_s' - \sigma^2 u_t]\boldsymbol{t} + [(u_s' + \kappa_g u_t)' + \tau_g(\kappa_n u_t - \tau_g u_s)]\boldsymbol{\nu}_{\mathcal{S}}\\ &\quad + [(\kappa_n u_t - \tau_g u_s)' - \tau_g(u_s' + \kappa_g u_t)]\boldsymbol{\nu},\end{aligned} \tag{A.5}$$

whence, after straightforward computations also involving differentiation of the identity (3.13), we arrive at

$$\text{(A.6)}\quad \begin{aligned}\mathcal{I}_2 &:= \int_{\mathcal{C}} \boldsymbol{a}' \cdot \boldsymbol{a}' \mathrm{d}s = \int_{\mathcal{C}} [\sigma^4 + (\kappa_g')^2 + (\kappa_n')^2 + \sigma^2\tau_g^2 + 2\tau_g(\kappa_n\kappa_g' - \kappa_g\kappa_n')]u_t^2 \\ &+ \int_{\mathcal{C}} (\sigma^2)' u_t u_t' + \int_{\mathcal{C}} \sigma^2 (u_t')^2 \mathrm{d}s + \int_{\mathcal{C}} 2[\kappa_g\sigma^2 - 2\tau_g(\kappa_n' - \tau_g\kappa_g)]u_t u_s' \\ &- \int_{\mathcal{C}} 2[\kappa_n\tau_g\sigma^2 + \tau_g^2(\kappa_g' + \kappa_n\tau_g) + \tau_g'(\kappa_n' - \tau_g\kappa_g)]u_t u_s - \int_{\mathcal{C}} 2[\tau_g(\kappa_n\kappa_g + 2\tau_g')]u_s u_s' \\ &+ \int_{\mathcal{C}} (\kappa_g^2 + 4\tau_g^2)(u_s')^2 + \int_{\mathcal{C}} [(\kappa_n\tau_g)^2 + \tau_g^4 + (\tau_g')^2]u_s^2 + \int_{\mathcal{C}} (u_s'')^2 + \int_{\mathcal{C}} 2\kappa_g u_t' u_s'' \\ &- \int_{\mathcal{C}} 2[\kappa_g\tau_g^2 + \tau_g'\kappa_n]u_s u_t' + \int_{\mathcal{C}} 2[\kappa_g' + \kappa_n\tau_g]u_t u_s'' - \int_{\mathcal{C}} 2\tau_g^2 u_s u_s'' - \int_{\mathcal{C}} 4\tau_g\kappa_n u_s' u_t' \,.\end{aligned}$$

To proceed, we consider the terms

$$\mathcal{I}_3 := \int_{\mathcal{C}} \sigma^2[(\boldsymbol{u}' \cdot \boldsymbol{t})^2 - \frac{3}{2}a^2]\mathrm{d}s,$$

which, by use of (A.3)–(A.4), can be recast as

$$\text{(A.7)}\quad \begin{aligned}\mathcal{I}_3 &= \int_{\mathcal{C}} \sigma^2 (u_t')^2 + \int_{\mathcal{C}} \sigma^2 \left(\kappa_g^2 - \frac{3}{2}\tau_g^2\right) u_s^2 - \int_{\mathcal{C}} 2\kappa_g\sigma^2 u_s u_t' - \int_{\mathcal{C}} \frac{3}{2}\sigma^4 u_t^2 - \int_{\mathcal{C}} \frac{3}{2}\sigma^2 (u_s')^2 - \\ &- \int_{\mathcal{C}} 3\sigma^2\kappa_g u_t u_s' + \int_{\mathcal{C}} 3\kappa_n\tau_g\sigma^2 u_t u_s \,.\end{aligned}$$

Finally, by integration by parts we change

$$\mathcal{I}_4 := -2\int_{\mathcal{C}} [(\boldsymbol{u}' \cdot \boldsymbol{t})'\sigma\boldsymbol{n} \cdot \boldsymbol{a} + 2(\boldsymbol{u}' \cdot \boldsymbol{t})\boldsymbol{a}' \cdot \sigma\boldsymbol{n}]\mathrm{d}s$$

into

$$\mathcal{I}_4 = 2\int_{\mathcal{C}} \boldsymbol{a} \cdot \sigma\boldsymbol{n}(\boldsymbol{u}' \cdot \boldsymbol{t})'\mathrm{d}s + 4\int_{\mathcal{C}} \boldsymbol{a} \cdot (\sigma\boldsymbol{n})'(\boldsymbol{u}' \cdot \boldsymbol{t})\mathrm{d}s,$$

which, also by use of $(3.4)_1$, (3.13), and (A.3), yields

$$\mathcal{I}_4 = \int_{\mathcal{C}} 2\sigma^2 u_t u_t'' + \int_{\mathcal{C}} 2\kappa_g u_s' u_t'' - \int_{\mathcal{C}} 2\kappa_n\tau_g u_s u_t'' + \int_{\mathcal{C}} 4[(\kappa_g' + \kappa_n\tau_g)]u_t' u_s'$$

$$\text{(A.8)}\quad \begin{aligned}&-2\int_{\mathcal{C}} \kappa_g^2 (u_s')^2 - \int_{\mathcal{C}} 2[\kappa_g[\kappa_n\tau_g + 3\kappa_g']]u_s u_s' + \int_{\mathcal{C}} 2[\kappa_g'\kappa_n\tau_g - 2\kappa_g(\kappa_g\tau_g^2 - \tau_g\kappa_n')]u_s^2 \\ &- \int_{\mathcal{C}} 2(\sigma^2\kappa_g)' u_s u_t + \int_{\mathcal{C}} 2(\sigma^2)' u_t u_t' + \int_{\mathcal{C}} 4(\kappa_g\tau_g - \kappa_n')\tau_g u_t' u_s - 2\int_{\mathcal{C}} \kappa_g\sigma^2 u_t u_s' \,.\end{aligned}$$

As already mentioned for $\delta\mathcal{F}^*$, the component u_t cannot appear in the final expression of $\delta^2\mathcal{F}^*$. To prove this, we collect terms in $\delta^2\mathcal{F}^*$ with the same dependence on u_t and its derivatives with respect to s, and integrate repeatedly by parts.

- Terms containing $u_t'^2$ and $u_t u_t''$ in $\mathcal{I}_3$ and $\mathcal{I}_4$ are

$$2\int_{\mathcal{C}} \sigma^2(u_t'^2 + u_t u_t'')\mathrm{d}s = -2\int_{\mathcal{C}}(\sigma^2)' u_t u_t' \mathrm{d}s, \tag{A.9}$$

which are combined with the term $3\int_{\mathcal{C}}\sigma^2 u_t u_t'\mathrm{d}s$ found in $\mathcal{I}_2$ and $\mathcal{I}_4$ to obtain, by (3.13),

$$\int_{\mathcal{C}}(\sigma^2)' u_t u_t'\mathrm{d}s = -\frac{1}{2}\int_{\mathcal{C}}(\sigma^2)'' u_t^2\mathrm{d}s = -\int_{\mathcal{C}}(\kappa_g'^2 + \kappa_n'^2 + \kappa_g\kappa_g'' + \kappa_n\kappa_n'')u_t^2\mathrm{d}s\,. \tag{A.10}$$

- Further terms containing u_t^2 in $\mathcal{I}_1$–$\mathcal{I}_3$ are collected together to yield, also by use of (3.13),

$$\int_{\mathcal{C}}\left[-\frac{\sigma^4}{2} + (\kappa_g')^2 + (\kappa_n')^2 + \tau_g^2\kappa_g^2 - 2\tau_g\kappa_g\kappa_n' + \frac{\sigma^2}{2}\kappa_n^2 + \kappa_n\kappa_n'' - \kappa_g\kappa_n\tau_g'\right]u_t^2\mathrm{d}s,$$

which, when added to (A.10), gives

$$\int_{\mathcal{C}}\left[-\frac{\sigma^4}{2} + \frac{\kappa_g}{2}(2\kappa_g\tau_g^2 - 4\tau_g\kappa_n' - 2\kappa_g'' - 2\kappa_n\tau_g') + \frac{\sigma^2}{2}\kappa_n^2\right]u_t^2\mathrm{d}s\,. \tag{A.11}$$

Now, terms in u_t should simplify separately for each integral in $\mathcal{F}$. Hence, we can use the reduced equilibrium equation

$$\kappa_g\sigma^2 + 2(\tau_g'\kappa_n + 2\tau_g\kappa_n' + \kappa_g'' - \kappa_g\tau_g^2) = 0, \tag{A.12}$$

obtained by setting $\delta\mathcal{F}^* = 0$ together with (3.13) to show that the integral (A.11) vanishes identically on any equilibrium configuration.

We now prove that mixed terms containing products of u_t and u_s or of their derivatives do not enter into $\delta^2\mathcal{F}^*$.

- The terms in $\mathcal{I}_2$ containing the product $u_t u_s''$ can be transformed via integration by parts as

$$2\int_{\mathcal{C}}(\kappa_g' + \kappa_n\tau_g)u_t u_s''\mathrm{d}s = -2\int_{\mathcal{C}}(\kappa_g' + \kappa_n\tau_g)u_t' u_s'\mathrm{d}s - 2\int_{\mathcal{C}} u_t u_s'(\kappa_g'' + \kappa_n'\tau_g + \kappa_n\tau_g')\mathrm{d}s,$$

to which we add the term in $\mathcal{I}_4$

$$2\int_{\mathcal{C}}\kappa_g u_t'' u_s'\mathrm{d}s = -2\int_{\mathcal{C}}(\kappa_g u_s'' u_t' + \kappa_g' u_s' u_t')\mathrm{d}s,$$

containing $u_t'' u_s'$ and then add the term in $\mathcal{I}_2$ that contains $u_s'' u_t'$: as a result, we are left with

$$-2\int_{\mathcal{C}}(2\kappa_g' + \kappa_n\tau_g)u_t' u_s'\mathrm{d}s - 2\int_{\mathcal{C}} u_t u_s'(\kappa_g'' + \kappa_n'\tau_g + \kappa_n\tau_g')\mathrm{d}s\,. \tag{A.13}$$

Since the remaining terms in $\mathcal{I}_2$ and $\mathcal{I}_4$ containing $u_t' u_s'$ reduce to

$$4\int_{\mathcal{C}}\kappa_g' u_t' u_s'\mathrm{d}s,$$

we finally arrive at

$$-2\int_{\mathcal{C}}\{[\kappa_g''+(\tau_g\kappa_n)']u_tu_s'+\kappa_n\tau_g u_t'u_s'\}\mathrm{d}s\,. \tag{A.14}$$

- The term

$$-2\int_{\mathcal{C}}\kappa_n\tau_g u_s u_t''\mathrm{d}s=2\int_{\mathcal{C}}[\kappa_n\tau_g u_t'u_s'+(\kappa_n\tau_g)'u_su_t']\mathrm{d}s$$

of $\mathcal{I}_4$ can be added to the integral (A.14) to arrive at

$$2\int_{\mathcal{C}}\{u_su_t'(\kappa_n\tau_g)'-[\kappa_g''+(\kappa_n\tau_g)']u_tu_s'\}\mathrm{d}s, \tag{A.15}$$

or, after integration by parts, at

$$-2\int_{\mathcal{C}}\{u_su_t(\kappa_n\tau_g)''+u_tu_s'[\kappa_g''+2(\kappa_n\tau_g)']\}\mathrm{d}s\,. \tag{A.16}$$

- We then consider the following terms in $\mathcal{I}_2$–$\mathcal{I}_4$ that contain $u_t'u_s$:

$$2\int_{\mathcal{C}}u_t'u_s[\kappa_g\tau_g^2-\tau_g'\kappa_n-\kappa_g\sigma^2-2\tau_g\kappa_n']\mathrm{d}s.$$

We integrate them by parts, obtaining

$$2\int_{\mathcal{C}}\{u_s'u_t[-\kappa_g\tau_g^2+\tau_g'\kappa_n+\kappa_g\sigma^2+2\tau_g\kappa_n']+u_tu_s[\tau_g'\kappa_n+\kappa_g\sigma^2+2\tau_g\kappa_n'-\kappa_g\tau_g^2]'\}\mathrm{d}s\,. \tag{A.17}$$

- Further integrals containing $u_s'u_t$ in $\mathcal{I}_2$–$\mathcal{I}_4$ are collected to give

$$-\int_{\mathcal{C}}[3\kappa_g\sigma^2+4\tau_g(\kappa_n'-\tau_g\kappa_g)]u_s'u_t\mathrm{d}s, \tag{A.18}$$

which, after algebraic manipulations and the use of (A.12), when added to (A.16) and (A.17) yield

$$2\int_{\mathcal{C}}u_su_t\{[\tau_g'\kappa_n+\kappa_g\sigma^2+2\tau_g\kappa_n'-\kappa_g\tau_g^2]'-(\kappa_n\tau_g)''\}\mathrm{d}s.$$

This simplifies to zero when it is added to the remaining terms in $\mathcal{I}_1$–$\mathcal{I}_4$ containing u_su_t, namely,

$$2\int_{\mathcal{C}}u_su_t[\tau_g^2\kappa_g'+2\tau_g\tau_g'\kappa_g-\tau_g'\kappa_n'-(\kappa_g\sigma^2)'-\kappa_n''\tau_g]\mathrm{d}s,$$

as can be easily checked.

Hence, we proved that only terms containing u_s'', u_s', and u_s appear in the second variation of $\mathcal{F}^*$. Precisely, we can recast $\delta^2\mathcal{F}^*$ into a diagonal form in which only $(u_s'')^2$, $(u_s')^2$, and $(u_s)^2$ appear.

- In $\mathcal{I}_2$ we consider the terms

$$\int_{\mathcal{C}}[(u_s'')^2-2\tau_g^2u_su_s'']\mathrm{d}s=\int_{\mathcal{C}}\{(u_s'')^2+2\tau_g^2(u_s')^2+4\tau_g\tau_g'u_su_s'\}\mathrm{d}s,$$

which, when added to the remaining terms in $\mathcal{I}_2$–$\mathcal{I}_4$ containing $(u_s')^2$, yield

$$\int_{\mathcal{C}}\left\{(u_s'')^2+\left(6\tau_g^2-\kappa_g^2-\frac{3}{2}\sigma^2\right)(u_s')^2+4\tau_g\tau_g' u_s u_s'\right\}\mathrm{d}s\,. \tag{A.19}$$

- We now add to (A.19) the terms in $\mathcal{I}_2$ and $\mathcal{I}_4$ with $u_s u_s'$ arriving at

$$\int_{\mathcal{C}}\left\{(u_s'')^2+\left(6\tau_g^2-\kappa_g^2-\frac{3}{2}\sigma^2\right)(u_s')^2-2[2\tau_g\kappa_g\kappa_n+3\kappa_g\kappa_g']u_s u_s'\right\}\mathrm{d}s\,. \tag{A.20}$$

Finally, if we add the remaining contributions in $\mathcal{I}_1$–$\mathcal{I}_4$ that contain u_s^2 and then integrate by parts the last term in (A.20), we obtain the expression (3.15) for $\delta^2\mathcal{F}^*$.

Acknowledgment. It is a pleasure to thank Prof. Epifanio G. Virga for his comments and suggestions.

REFERENCES

[1] J. W. Gibbs, *On the equilibrium of heterogeneous substances*, in The Collected Papers of J. Willard Gibbs, Vol. I, Yale University Press, London, 1957, pp. 55–353.

[2] R. Tolman, *The effect of droplet size on surface tension*, J. Chem. Phys., 17 (1949), pp. 333–337.

[3] M. J. P. Nijmeijer, C. Bruin, A. B. Van Woerkom, A. F. Bakker, and J. M. J. Van Leeuwen, *Molecular dynamics of the surface tension of a drop*, J. Chem. Phys., 96 (1992), pp. 565–576.

[4] E. M. Blockhuis and D. Bedeaux, *Pressure tensor of a spherical interface*, J. Chem. Phys., 97 (1992), pp. 3576–3586.

[5] E. M. Blockhuis and D. Bedeaux, *Derivation of microscopic expressions for the rigidity constants of a simple liquid vapor interface*, Phys. A, 184 (1992), pp. 42–70.

[6] V. G. Baidakov, G. Sh. Boltachev, and G. G. Chernykh, *Curvature corrections to surface tension*, Phys. Rev. E (3), 70 (2004), 011603.

[7] P. S. Swain and R. Lipowsky, *Contact angles on heterogeneous substrates: A new look at Cassie's and Wenzel's laws*, Langmuir, 14 (1998), pp. 6772–6780.

[8] R. Lipowsky, P. Lenz, and P. S. Swain, *Wetting and dewetting of structured and imprinted surfaces*, Colloids Surf. A, 161 (2000), pp. 3–22.

[9] D. J. Steigmann and D. Li, *Energy-minimizing states of capillary systems with bulk, surface, and line phases*, IMA J. Appl. Math., 55 (1995), pp. 1–17.

[10] Y. Solomentsev and L. R. White, *Microscopic drop profiles and the origins of line tension*, J. Colloid Interf. Sci., 218 (1999), pp. 122–136.

[11] J. S. Rowlinson and B. Widom, *Molecular Theory of Capillarity*, Dover, New York, 2002.

[12] R. Rosso and E. G. Virga, *A general stability criterion for wetting*, Phys. Rev. E (3), 68 (2003), 012601.

[13] R. Rosso and E. G. Virga, *Local stability for a general wetting functional*, J. Phys. A, 37 (2004), pp. 3989–4015.

[14] R. Rosso and E. G. Virga, *Corrigendum: Local stability for a general wetting functional*, J. Phys. A, 37 (2004), p. 8751.

[15] M. Brinkmann, J. Kierfeld, and R. Lipowsky, *A general stability criterion for droplets on structured substrates*, J. Phys. A, 37 (2004), pp. 11547–11573.

[16] R. Rosso and E. G. Virga, *Sign of line tension in liquid bridge stability*, Phys. Rev. E (3), 70 (2004), 031603.

[17] M. Brinkmann, J. Kierfeld, and R. Lipowsky, *Stability of liquid channels or filaments in the presence of line tension*, J. Phys. Condens. Matter, 17 (2005), pp. 2349–2364.

[18] L. Guzzardi, R. Rosso, and E. G. Virga, *Residual stability of sessile droplets with negative line tension*, Phys. Rev. E (3), 73 (2006), 021602.

[19] L. Guzzardi and R. Rosso, *Sessile droplets on a curved substrate: Effects of line tension*, J. Phys. A, 40 (2007), pp. 19–46.

[20] J. Y. Wang, S. Betelu, and M. Law, *Line tension approaching a first-order wetting transition: Experimental results from contact angle measurements*, Phys. Rev. E (3), 63 (2001), 031601.

[21] L. Boruvka and A. W. Neumann, *Generalization of the classical theory of capillarity*, J. Chem. Phys., 66 (1977), pp. 5464–5476.

[22] A. I. Rusanov, A. K. Shchekin, and D. V. Tatyanenko, *The line tension and the generalized Young equation: The choice of dividing surface*, Colloids Surf. A, 250 (2004), pp. 263–268.

[23] P. Jakubczyk and M. Napiórkowski, *Influence of inhomogeneous substrate curvature on line tension*, Phys. Rev. E (3), 72 (2005), 011603.

[24] T. Bieker and S. Dietrich, *Wetting of curved substrates*, Phys. A, 252 (1998), pp. 85–137.

[25] R. Rosso, *Curvature effects in vesicle-particle interactions*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 829–852.

[26] M. P. Do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice–Hall, Englewood Cliffs, NJ, 1976.

[27] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers. Asymptotic Methods and Perturbation Theory*, Springer, Heidelberg, Germany, 1999.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 552–564

STATISTICS OF POLARIZATION-MODE DISPERSION EMULATORS WITH UNEQUAL SECTIONS*

BRENTON R. STONE[†], GINO BIONDINI[‡], AND WILLIAM L. KATH[§]

Abstract. We study two models for the generation of polarization-mode dispersion (PMD) with unequal, fixed-length sections: an isotropic model, in which the orientations of all the sectional PMD vectors are taken to be randomly and uniformly varying across the Poincaré sphere, and a rotator model, in which all sections are taken to be linearly birefringent waveplates randomly rotatable with respect to one another. We describe the implementation of importance sampling for first- and second-order PMD in both models, including a targeting method for first-order PMD. We then use analytical and numerical methods to reconstruct the statistics of first- and second-order PMD for the two models. Our results show that the statistical properties of PMD depend significantly on the specific details of how PMD is generated.

Key words. optical fiber communications, Monte Carlo methods, importance sampling

AMS subject classifications. 65C05, 65C30, 78A10, 78A40, 78A48, 90B18

DOI. 10.1137/070696350

1. Introduction.

Polarization-mode dispersion (PMD) is one of the major challenges facing the next generation of optical fiber communication systems [20]. Optical fiber is slightly birefringent due to slight deviations from circular symmetry, bending, stresses, etc. To first order in frequency, birefringence splits a pulse between the fast and the slow axes in an optical fiber; higher orders of birefringence induce depolarization and polarization-dependent chromatic dispersion. Also, the birefringence properties change randomly with distance, temperature, time, and wavelength, and these random variations are referred to as PMD. In system design, a certain power penalty is usually allotted to PMD, and one demands that the outage probability (the probability of the PMD-induced penalty exceeding this allowed value) be very small (typical requirements are a minute per year). Because of this stringent requirement, it has been difficult to use either Monte Carlo (MC) simulations or laboratory measurements to fully assess system outage probabilities, due to the extremely large number of PMD configurations that are necessary to obtain reliable estimates. Recently, it was shown that the technique of importance sampling (IS) [7, 8, 11] can often obviate this problem and allow efficient computation of PMD-induced transmission penalties and outage probabilities [6, 24, 25].

A measure of PMD is provided by the PMD vector [16]. The magnitude of the PMD vector, called differential group delay (DGD), quantifies the amount of local pulse splitting between fast and slow axes of birefringence. It has long been assumed that the probability density function (PDF) of the DGD follows a Maxwellian distribution [12], and that the process is ergodic, in the sense that time averages coincide

[*] Received by the editors July 5, 2007; accepted for publication (in revised form) July 29, 2008; published electronically December 3, 2008. This work was partially supported by the National Science Foundation under grants DMS-0406513 and DMS-0506101.

http://www.siam.org/journals/siap/69-2/69635.html

[†] Department of Mathematics, State University of New York, Buffalo, NY 14260 (brenton.stone@tyndall.af.mil). Current address: Applied Research Associates, Tyndall AFB, Panama City, FL 32403.

[‡] Department of Mathematics, State University of New York, Buffalo, NY 14260 (biondini@buffalo.edu).

[§] Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (kath@northwestern.edu).

with frequency averages. Recent measurements of installed fiber links, however, have reported variations in the temporal statistics of DGD between different frequency channels in wavelength-division-multiplexed (WDM) systems [9, 19]. This more complicated behavior is consistent with a so-called hinge model of PMD [9], in which the system is composed of a concatenation of a small number of long, stable sections (long stretches of fiber which are buried underground) joined by short, unprotected sections, or "hinges" (bridges, amplifiers, service huts, etc.), which are more subject to environmental effects. While the hinges themselves bring little or no contribution to the total DGD of the system, their random fluctuations appear to be responsible for the temporal dynamics of PMD within each channel, whereas the longer sections appear to be essentially frozen in time. In the traditional model of PMD (which can be thought of as the limit in which the number of hinges is large and the stable sections are short), different wavelength bands behave independently but share the same statistical properties. In contrast, in the hinge model different wavelength bands are not statistically identical (because the individual PMD vector of each section is different for each wavelength). Thus, the ergodic hypothesis is not satisfied in the hinge model.

The properties of PMD in the hinge model have been well characterized [1, 2, 3, 10, 17, 21, 23, 26] under the assumption that the hinges randomize the orientation of the PMD vectors uniformly across the Poincaré sphere. In particular, analytical expressions for the PDF of the DGD are available [1, 17]. No analytical expressions exist, however, for the PDF of higher-order PMD, including second-order PMD. Moreover, some of the features of the hinge model also apply for different mechanisms of PMD generation that give rise to concatenations of fixed-length sections. No analytical expressions, however, are known for the statistics of PMD if one relaxes the assumption that the individual sections are uniformly distributed on the Poincaré sphere.

Here we address both of the above issues, and we discuss the PMD statistics produced by a finite concatenation of fixed-length sections for two specific PMD generation models. In the first model we take the orientation of the individual sections to be uniformly distributed on the Poincaré sphere. In the second model we take the individual sections to be linearly birefringent, with randomly oriented axes with respect to one another. We refer to the first and second models, respectively, as the "isotropic" and the "rotator" models of PMD. For both models, we study the case in which the individual section lengths are not all identical. Previously, we discussed the implementation of IS techniques for both models in the special case of equal-length sections [7, 8, 11]. Here we extend those results to the case of nonequal lengths, discussing the generation of large values of DGD, second-order PMD, as well as any combination of the two, plus a targeting method that allows one to concentrate samples where desired. Finally, we apply these methods to reconstruct PMD statistics of both models and show that significant differences exist among them.

2. Isotropic and rotator PMD models. The action of any lossless transmission element on an optical pulse can be described, up to a polarization-independent factor, by a unitary 2×2 frequency-dependent transmission matrix $U(z, \omega)$ called the Jones matrix, which describes the evolution of the transverse components of the optical field. Polarization effects can then be uniquely characterized by the real three-component PMD vector, $\vec{\tau}(\omega, z)$, defined by [16]

$$\vec{\tau}(\omega, z) \cdot \vec{\sigma} = 2i \frac{\partial U}{\partial \omega} U^{-1} , \tag{2.1}$$

where $\vec{\sigma}$ is a vector of Pauli matrices. Consider a concatenation a finite number of fixed-length fiber sections. The growth of PMD is governed, at each frequency, by the PMD concatenation equations. For first and second order, these are [16]

$$\vec{\tau}^{(n+1)} = \mathsf{R}_{n+1}\vec{\tau}^{(n)} + \Delta\vec{\tau}_{n+1}, \tag{2.2a}$$

$$\vec{\tau}_{\omega}^{(n+1)} = \mathsf{R}_{n+1}\vec{\tau}_{\omega}^{(n)} + \Delta\vec{\tau}_{n+1} \times \vec{\tau}^{(n+1)} + \Delta\vec{\tau}_{\omega,n+1}\,. \tag{2.2b}$$

Here $\vec{\tau}^{(n)}$ is the total PMD vector after the nth section, the fixed vector $\Delta\vec{\tau}_n$ is the PMD vector of the nth section, $\Delta\vec{\tau}_{\omega,n}$ is its frequency derivative, and the 3×3 matrix R_n is the Müller matrix of the nth section, which is related to the Jones matrix of that section by [16] $\mathsf{R}(\omega,z)\,\vec{\sigma} = U^{-1}\vec{\sigma}U$.

As is customary in both the traditional and the hinge model of PMD, we assume the sectional PMD vectors $\Delta\vec{\tau}_n$ to be constant in time and to have independent, identically distributed components that follow a normal distribution with mean zero and variance σ^2 with respect to wavelength. This of course implies that the sectional DGDs $\Delta\tau_n = |\Delta\vec{\tau}_n|$ are Maxwellian-distributed with respect to wavelength. Moreover, we assume that each section is linearly birefringent in frequency, namely, $\Delta\vec{\tau}_{\omega,n} = 0$. The matrix R_n then describes a rotation about an angle ϕ_n about the axis $\hat{r}_n = \Delta\vec{\tau}_n/|\Delta\vec{\tau}_n|$, namely,

$$\mathsf{R}_n = \exp\left[\phi_n\,\hat{r}_n \times \quad\right] = \cos\phi_n\,\mathsf{I}_3 + (1-\cos\phi_n)\,\hat{r}_n\hat{r}_n^T + \sin\phi_n\,\hat{r}_n \times \quad , \tag{2.3}$$

where I_3 is the 3×3 identity matrix and the superscript T denotes the matrix transpose. If hinges are present, a hinge rotation matrix H_n precedes R_{n+1} in (2.2).

In other words, mathematically there are two distinct random processes taking place. The first one governs the selection of the fiber sections, resulting in a set of wavelength-dependent sectional PMD vectors. These PMD vectors are stable over long periods of time (often months). The second random process is the one that governs the fast temporal variations due to environmental effects, and affecting the rotation angles in the matrices (2.3) as well as H_n. We model this situation by taking the section lengths to be fixed and by assuming that the only temporal variation in (2.2) arises from the rotation matrices R_{n+1} and, if present, from H_n. If the action of the hinges is sufficient to scatter the orientation of the PMD vectors uniformly across the Poincaré sphere, it is convenient to rewrite (2.2) as

$$\vec{\tau}^{(n+1)} = \mathsf{R}_{n+1}\mathsf{H}_n\big(\vec{\tau}^{(n)} + \Delta\vec{\tau}\,'_{n+1}\big), \tag{2.4a}$$

$$\vec{\tau}_{\omega}^{(n+1)} = \mathsf{R}_{n+1}\mathsf{H}_n\big(\vec{\tau}_{\omega}^{(n)} + \Delta\vec{\tau}\,'_{n+1} \times \vec{\tau}^{(n)}\big)\,, \tag{2.4b}$$

where $\Delta\vec{\tau}\,'_{n+1} = (\mathsf{R}_{n+1}\mathsf{H}_n)^{-1}\Delta\vec{\tau}_{n+1}$ is now uniformly distributed across the Poincaré sphere. We refer to this as the "isotropic" model of PMD. If one is interested only in first-order PMD, the problem is then equivalent to a 3-dimensional random walk. The DGD in the isotropic model and its impact on system behavior has been well characterized [1, 2, 3, 21, 26]. The isotropic hypothesis is a convenient assumption, since it makes an analytical treatment of the model possible. We emphasize, however, that it has not been experimentally validated, and that the question of whether or not it is accurate is an open issue.

When PMD is generated by birefringent waveplates, the individual PMD vectors lie on the equatorial plane of the Poincaré sphere. Since optical fiber is naturally linearly birefringent [14], the same statement holds for short fiber sections. Therefore, here we also consider the case in which the vectors $\Delta\vec{\tau}_{n+1}$ in (2.2) are uniformly distributed on the equatorial plane of the Poincaré sphere. We refer to this as the "rotator" model of PMD. Of course, because the fiber birefringence axes change with distance, the concatenated total PMD vector will wander off the equatorial plane. Since the fiber correlation length (which is the distance over which the birefringence properties become uncorrelated) is below a hundred meters [15], any fiber span longer than a few kilometers will have a PMD vector that does not lie in the equatorial plane. (At the same time, however, significant nonuniformities in the angular distribution of PMD persist up to medium-to-long distances [28].) Note also that the rotator model neglects the action of the hinge rotation matrix. Thus it is an oversimplification of the actual PMD generation mechanism in installed systems.

Even though both of the above-mentioned fixed-length models may not be a fully accurate representation of the actual mechanism of PMD generation, in the absence of more complete models or conclusive experimental data the comparison between them will serve to demonstrate that the statistical properties of PMD depend significantly on the physical details of how PMD is generated in the system. (These results generalize those previously obtained for the case of equal-length sections [7, 11].)

The case of unequal-length sections has recently received renewed interest (e.g., see [17, 18, 26]), but only within the framework of the isotropic assumption. The case of unequal lengths is worthy of study because, while no analytical expressions exist for the PDF of the DGD (for the rotator model) or second-order PMD (for either model), most installed systems are composed of sections with unequal lengths. We also emphasize that a key assumption in both of the models considered here is that the individual sections have *fixed length*, namely, that the sectional PMD vectors are essentially frozen. A different model, in which the individual PMD vectors are also varying and in which, in particular, they are Maxwellian-distributed, was studied in [5, 22]. Note, however, that allowing the section lengths to vary on the same temporal scales as the rotation matrices results in very different PMD properties from those of the models considered here, even in the isotropic case (e.g., the PDF of the total DGD is exactly Maxwellian-distributed for any number of sections).

3. Importance sampling for unequal sections.

Here we extend the IS methods that were derived in [7, 8, 11] for equal-length sections to the case of unequal section lengths, when the individual sections are either uniformly distributed on the Poincaré sphere or linearly birefringent. As mentioned previously, analytical expressions for the PDF of the DGD for the isotropic model are of course available both for equal and for unequal section lengths [1, 4, 17]. No similar expressions are known, however, when PMD is generated by birefringent waveplates. Moreover, no analytic expressions are known for the PDF of second-order PMD in either model. It was shown [24] that in systems which employ PMD compensation, knowledge of first-order PMD is not enough to accurately characterize PMD-induced transmission penalties. At the same time, it has also been shown that control of both first- and second-order PMD is enough in most cases of interest [24]. Finally, we should note that one of the advantages of IS is that, whether one is biasing for first- or second-order PMD, the method automatically generates PMD of all orders (due to the wavelength dependence of the sectional PMD vectors). This means that biasing for large DGDs can also be useful in the isotropic model (even though the PDF of the DGD is known), since the simultaneous presence of all orders of PMD can lead to a more accurate evaluation of PMD-induced distortions [24].

3.1. IS for the DGD. The first step when applying IS is to determine the most likely system configurations that lead to the event of interest. For first-order PMD, this proceeds exactly in the same way as in the case of equal-length sections [7]. Namely, one must choose the next PMD contribution $\Delta\vec{\tau}_{n+1}$ to be preferentially aligned with the previous PMD vector $\vec{\tau}^{(n)}$. For the isotropic model, this is done choosing the angle θ_n between $\Delta\vec{\tau}_{n+1}$ and $\vec{\tau}^{(n)}$ to be preferentially close to 0. For the rotator model, instead, since $\Delta\vec{\tau}_{n+1}$ must lie in the equatorial plane of the Poincaré sphere, the biasing is done by choosing the angle θ_n between $\Delta\vec{\tau}_{n+1}$ and the projection of $\vec{\tau}^{(n)}$ onto the equatorial plane to be preferentially close to 0.

To achieve this preferential alignment, for the isotropic model we take $\cos\theta = 2x^{1/\alpha} - 1$, while for the rotator model we take $\theta = \pi\ \mathrm{sgn}(2x-1)|2x-1|^\alpha$, in both cases with x uniform in $[0,1]$. These choices correspond respectively to the biasing distributions

$$p_{\alpha,\mathrm{iso}}(\theta) = (\alpha/2)\sin\theta\,[(1+\cos\theta)/2]^{\alpha-1}, \qquad p_{\alpha,\mathrm{wav}}(\theta) = (1/\alpha\pi)\,|\theta/\pi|^{1-\alpha}. \tag{3.1}$$

For both the isotropic and the rotator models the value $\alpha = 1$ reproduces the unbiased case, while larger values of α concentrate the samples θ near 0. In both models, the likelihood ratio is given by [8]

$$L(\theta_1,\dots,\theta_N) = \prod_{n=1}^{N} \frac{p_1(\theta_n)}{p_\alpha(\theta_n)}\,, \tag{3.2}$$

where N is the total number of sections. Of course, other choices of biasing distributions might work equally well, as long as the reference angles are correctly identified. Also, in both models the rotation angle ϕ_n is not important for IS purposes, and is taken to be varying and uniformly distributed in $[0,2\pi]$.

3.2. IS for first- and second-order PMD. Consider the orthogonal frame of reference defined by the unit vectors

$$\hat{u}_1^{(n)} = \vec{\tau}^{(n)}/|\vec{\tau}^{(n)}|\,, \qquad \hat{u}_2^{(n)} = \vec{\tau}_{\omega,\perp}^{(n)}/|\vec{\tau}_{\omega,\perp}^{(n)}|\,, \qquad \hat{u}_3^{(n)} = \hat{u}_1^{(n)} \times \hat{u}_2^{(n)}\,, \tag{3.3}$$

where $\vec{\tau}_{\omega,\|}$ and $\vec{\tau}_{\omega,\perp}$ are, respectively, the parallel and perpendicular components of $\vec{\tau}_\omega$ with respect to τ. As in [8], it is convenient to consider the continuum limit of (2.2). In the isotropic model, factoring out the inessential rotation $\mathsf{R}_{n+1}\mathsf{H}_n$ in (2.4), one then obtains

$$\frac{d\tau}{dz} = b_1\,, \qquad \frac{d\tau_{\omega,\|}}{dz} = b_2\,\frac{\tau_{\omega,\perp}}{\tau}\,, \qquad \frac{d\tau_{\omega,\perp}}{dz} = b_3\tau - b_2\,\frac{\tau_{\omega,\|}}{\tau}\,, \tag{3.4}$$

where $\vec{b}(z) = \lim_{\Delta z\to 0}\Delta\vec{\tau}_{n+1}/\Delta z$ quantifies the rate at which PMD is added, and (b_1,b_2,b_3) are the components of $\vec{b}$ with respect to the reference frame $\{\hat{u}_1,\hat{u}_2,\hat{u}_3\}$.

As shown in [8], (3.4) can be solved exactly for any $\vec{b}(z)$. One can then use calculus of variations to find the choice of $\vec{b}(z)$ which maximizes second-order PMD. In the case of equal-length sections, that is, $|\vec{b}(z)| = b = \text{const}$, it was shown in [11] that the solution of this maximization problem is

$$\vec{b}(z) = b\left(\hat{u}_1\,\cos[\Phi(z)] + \hat{u}_3\,\sin[\Phi(z)]\right), \tag{3.5}$$

where the biasing angle is $\Phi(z) = (z/z_{\max})\Phi_{\max}$ and where $\Phi_{\max} = \pi/2$. Other choices of $\Phi_{\max}$ maximize linear combinations of DGD and second-order PMD ($\Phi_{\max} = 0$ is simply first-order biasing).

As we show next, the solution of the maximization problem in the case of unequal section lengths can be obtained from the case of equal section lengths by a simple change of variable. Given a function $b(z) = |\vec{b}(z)|$ that describes the magnitude of the local birefringence, define the rescaled distance

$$\zeta(z) = \int_0^z b(z')dz' \,. \tag{3.6}$$

Written in terms of ζ, equations (3.4) then are

$$\frac{d\tau}{d\zeta} = e_1 \,, \qquad \frac{d\tau_{\omega,||}}{d\zeta} = e_2 \,\frac{\tau_{\omega,\perp}}{\tau} \,, \qquad \frac{d\tau_{\omega,\perp}}{d\zeta} = e_3\tau - e_2 \,\frac{\tau_{\omega,||}}{\tau} \,, \tag{3.7}$$

where $\hat{e}(z) = \vec{b}(z)/|\vec{b}(z)|$ has unit magnitude. One can now use the results of the calculations for equal-length sections described above, obtaining

$$\hat{e}(\zeta) = \hat{u}_1 \, \cos[\Phi(\zeta)] + \hat{u}_3 \, \sin[\Phi(\zeta)] \,, \tag{3.8}$$

with $\Phi(\zeta) = (\zeta/\zeta_{\max})\,\Phi_{\max}$. In terms of the original variable z one then obtains (3.5), with b replaced by $b(z)$ and where now

$$\Phi(z) = \Phi_{\max} \int_0^z b(z')dz' \Big/ \int_0^{z_{\max}} b(z')dz' \,, \tag{3.9}$$

with the same meaning for $\Phi_{\max}$. (Of course, (3.9) reduces to the linearly varying profile (3.10) in the case of equal section lengths.) In the discrete version, with section lengths $\Delta\tau_n$, we then obtain Φ_n, which gives the proper biasing direction after n sections as a function of n:

$$\Phi_n = \Phi_{\max} \sum_{m=1}^{n} \Delta\tau_m/\tau_{\max}, \tag{3.10}$$

where

$$\tau_{\max} = \sum_{n=1}^{N} \Delta\tau_n \,. \tag{3.11}$$

As with the DGD, the biasing directions for the rotator model are just the projection of the vector $\vec{b}$ onto the equatorial plane. Moreover, once the deterministic biasing directions have been found, for both the isotropic and the rotator model, one selects the biasing distributions in order to concentrate the MC samples around the deterministic biasing directions exactly as when biasing for the DGD. When more than one choice of biasing is necessary, multiple IS can be used, and samples from different biasing distributions can be combined using the balance heuristic [8].

3.3. Targeted IS. It is sometimes useful to be able to choose where to concentrate the MC samples, especially when one is interested in only a small region of the PDF. We now obtain an estimate of the values of the biasing strength α that are needed to generate values of DGD concentrated near a given target.

A simple bound on α can be obtained by looking at the component of the next PMD vector, $\tau^{(n+1)}$, that is parallel to the previous PMD vector, $\tau^{(n)}$. From (2.4a) it is

$$\tau_{||}^{(n+1)} = \tau^{(n)} + \Delta\tau_{n+1} \cos\theta_{n+1}, \tag{3.12}$$

where θ_{n+1} is the angle between the differential contribution of the next section, $\Delta\tau_{n+1}$, and the previous PMD vector. Since it is obviously true that $|\vec{\tau}^{(n)}| \geq |\tau_{||}^{(n)}|$, from (3.12) we have

$$\tau_{||}^{(N)} \geq \sum_{n=1}^{N} \Delta\tau_n \cos\theta_n . \tag{3.13}$$

Then, since the section lengths are fixed, taking expectation values we have

$$\langle \tau^{(N)} \rangle \geq \sum_{n=1}^{N} \Delta\tau_n \langle \cos\theta_n \rangle = \langle \cos\theta \rangle \, \tau_{\max} \, , \tag{3.14}$$

where we used $\langle \tau^{(N)} \rangle \geq \langle \tau_{||}^{(N)} \rangle$, and where $\tau_{\max}$ is the maximum DGD, defined in (3.11). For the biased probability distribution for the isotropic model, (3.1), it is

$$\langle \cos\theta \rangle = (\alpha - 1)/(\alpha + 1). \tag{3.15}$$

Therefore in this case (3.14) yields

$$\langle \tau^{(N)} \rangle \geq [(\alpha - 1)/(\alpha + 1)] \, \tau_{\max} \, . \tag{3.16}$$

Note that as α increases, $\langle \tau^{(N)} \rangle$ approaches $\tau_{\max}$, as expected.

Equation (3.16) yields an upper bound on the value of the biasing parameter α that is necessary to obtain samples distributed near $\langle \tau^{(N)} \rangle = \tau_{\text{target}}$, namely, $\alpha_{\text{target}} \leq \alpha_o$, where

$$\alpha_o = (\tau_{\max} + \tau_{\text{target}})/(\tau_{\max} - \tau_{\text{target}}) \, . \tag{3.17}$$

Note, however, that the above result is only an upper bound, and the actual value of α_{target} that concentrates the samples around a desired target is often a great deal smaller than α_o, particularly if $\tau_{\max} - \tau_{\text{target}}$ is small. In those cases, the heuristic correction $\alpha = \alpha_o/\sqrt{1 + \alpha_o/4}$, where α_o is given by (3.17), provides a better estimate of α_{target} which is valid over the whole range of DGDs.

The bound on α_{target} is independent of the particular values of the section lengths, and therefore it holds equally well for equal as well as unequal lengths. Moreover, it also appears to hold for the rotator model. It should be noted that for PMD generation models in which the section lengths are also variable and are Maxwellian-distributed, one could also concentrate the samples around a target value of second-order PMD [5]. Note also that, using the so-called Brownian bridge method, one could hit exactly any value of DGD [27]. To the best of our knowledge, however, no targeting method was known for fixed-length sections.

4. Numerical simulations and PMD statistics.

We now proceed to compare statistical measures of PMD for the isotropic and the rotator models of PMD defined in section 2, using the methods discussed in section 3.

4.1. PDF of the DGD.

An exact formula for the PDF of the DGD in the isotropic model, hereafter $p_{\text{DGD}}(\tau)$, was obtained in [1]. That expression, however, involves a sum over 2^N terms, where as before N is the number of sections. Hence, the computational cost of evaluating it increases exponentially with N. As a result, the use of that formula is impractical except for very small values of N. Fortunately, an alternative exact expression exists for the PDF of the DGD, in terms of a Fourier sine series [4]:

$$p_{\mathrm{DGD}}(\tau) = \frac{2\pi\tau}{\tau_{\max}^2} \sum_{m=1}^{\infty} m \sin\left(\frac{m\pi\tau}{\tau_{\max}}\right) \prod_{n=1}^{N} \frac{\sin(m\pi\Delta\tau_n/\tau_{\max})}{m\pi\Delta\tau_n/\tau_{\max}} \,. \tag{4.1}$$

The evaluation of (4.1) is of course always affected by truncation error due to the finite number of Fourier modes. Nonetheless, (4.1) produced the same results as the formula in [1] up to roundoff error in all the cases we tested (which included situations where the DGD of one section is larger than the sum of all the others). In our tests the two methods had about the same execution time for $N = 6$ sections when 2^{10} Fourier modes were used in (4.1). The computational cost of evaluating (4.1), however, is essentially given by the number of Fourier modes used, and depends only very weakly on the number of sections [23]. For this reason we used (4.1) in our calculations. For the rotator model one must use importance-sampled MC (IS-MC) simulations as discussed in section 3.1, using a combination of biasing strengths, to cover the whole range of DGDs. The PDF of the DGD is then reconstructed from the IS-MC simulations using the likelihood ratios as described in [8].

Figure 4.1 shows the PDF of the DGD for the isotropic model (thick dashed line) and the rotator model (thick solid line) for two particular realizations of $N = 8$ section lengths, while Figure 4.2 shows the PDF of the DGD for a specific realization of $N = 20$ sections. The specific values of the sectional DGDs are given in Table 4.1; the corresponding values of $\tau_{\max}$ are, respectively, 10.99 ps, 16.19 ps, and 21.60 ps for cases A, B, and C. In each case, the sectional DGDs were all drawn from an identical Maxwellian distribution

$$p_{\mathrm{dgd}}(\tau) = \frac{\sqrt{2}\,\tau^2}{\sqrt{\pi}\,\sigma^3}\, e^{-\tau^2/2\sigma^2} \,, \tag{4.2}$$

where $\sigma^2 = (\pi/8)\langle\Delta\tau\rangle^2$. In particular, we set $\langle\Delta\tau\rangle = \langle\tau\rangle/\sqrt{N}$ with $\langle\tau\rangle = 5\,\mathrm{ps}$ in all cases, so as to obtain a nominal Maxwellian distribution with mean DGD of 5 ps for the whole line. Note, however, that due to the finite sample size (i.e., the finite value of N), the samples will generate a PDF of the DGD which is better approximated by an "effective" Maxwellian, obtained by (4.2) with $\sigma_{\mathrm{eff}}^2 = (\pi/8)\sum_{n=1}^{N}(\Delta\tau_n)^2$. These effective Maxwellian distributions are shown in Figure 4.1 as dot-dashed lines. Of course the difference between the nominal and effective Maxwellians will tend to zero on average as N goes to infinity.

For the rotator model we used the biasing strengths $\alpha = 1$ (unbiased), 4, 12, and 24 to perform IS-MC simulations in the cases with $N = 8$, and we used $\alpha = 1$, 2, 4, and 6 in the case $N = 20$ (since in this case smaller biasing strengths are sufficient to cover the desired range of values of the PDF). In all cases 400,000 samples per biasing strength were used. Note how, in all cases, the tails of the PDF for the isotropic model (that is, the values of the PDF for values of DGD near $\tau_{\max}$) are orders of magnitudes below those of the rotator model. This behavior occurred in all cases we studied, but it is not clear at present whether it is a general property, namely, whether it holds for any choice of section lengths and for any number of sections (as is indeed the case with equal lengths). As is to be expected, the PDF in the case of $N = 20$ sections agrees with a Maxwellian distribution over a larger range of DGDs.

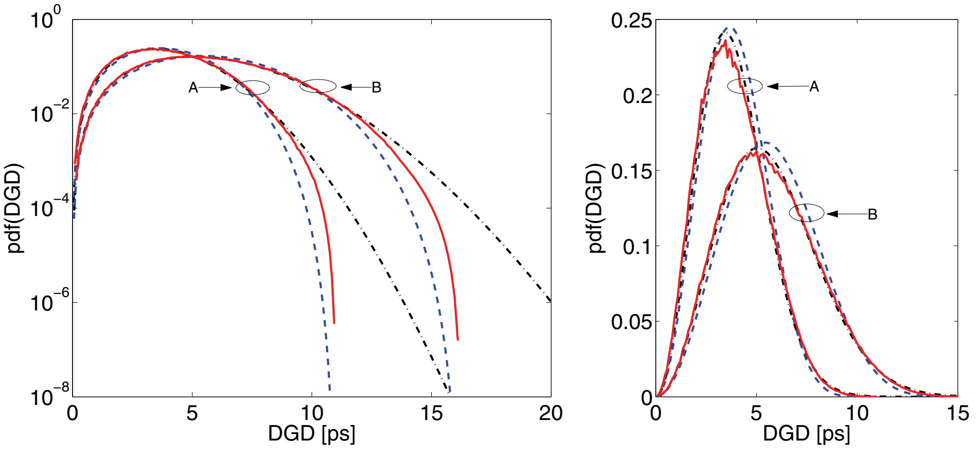


Fig. 4.1. *PDF of the DGD for a concatenation of $N = 8$ sections, for two particular choices of individual section DGDs drawn from an identical Maxwellian distribution with mean $5/\sqrt{N}$ ps (cases* A *and* B *in Table* 4.1*). Dashed lines: isotropic model; solid lines: rotator model. The dot-dashed lines show the effective Maxwellian distribution. Left: logarithmic scale; right: linear scale.*

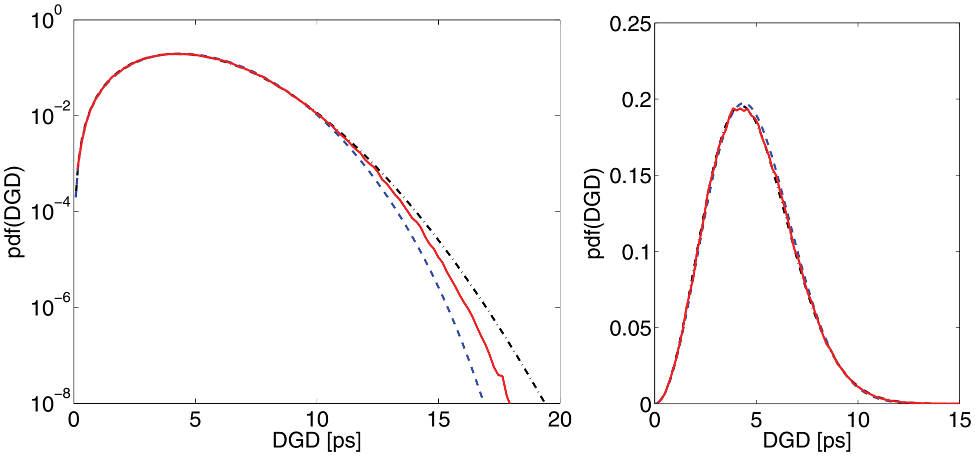


Fig. 4.2. *Same as Figure* 4.1*, but for a concatenation of $N = 20$ sections drawn from a Maxwellian distribution with mean $5/\sqrt{N}$ ps (case* C *in Table* 4.1*).*

Table 4.1

Sectional DGDs (in ps) used in the MC simulations.

A		1.149	2.077	1.390	2.094	1.260	0.2761	1.812	0.9307	
B		1.632	2.278	1.678	3.584	2.034	0.948	1.164	2.868	
C	1.170	1.624	0.554	1.127	1.232	0.450	0.916	0.589	1.094	1.236
	1.230	0.824	1.243	0.997	0.511	0.628	1.397	1.853	0.841	2.089

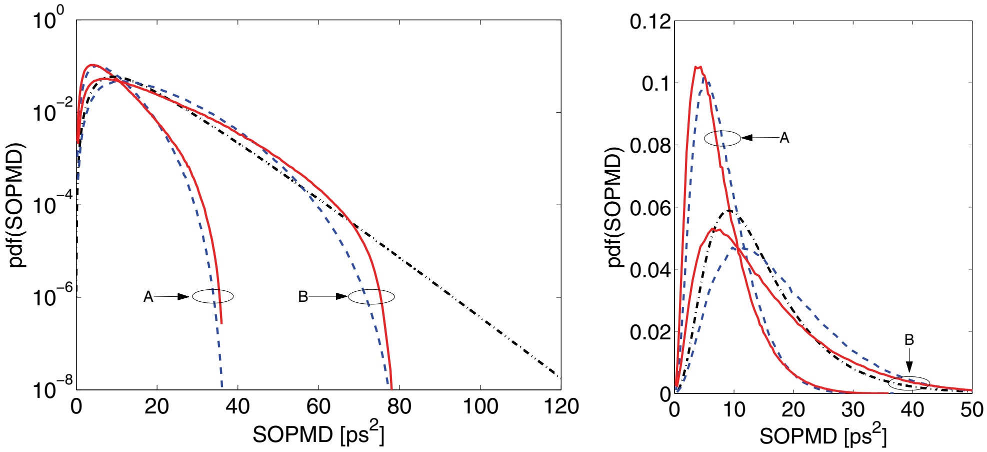


FIG. 4.3. *PDF of second-order PMD (SOPMD) for the same section DGDs as in Figure* 4.1. *Dashed lines: isotropic model; solid lines: rotator model. The dot-dashed line shows a sech-tanh distribution with* $\langle \tau \rangle = 5$ *ps. Left: logarithmic scale; right: linear scale.*

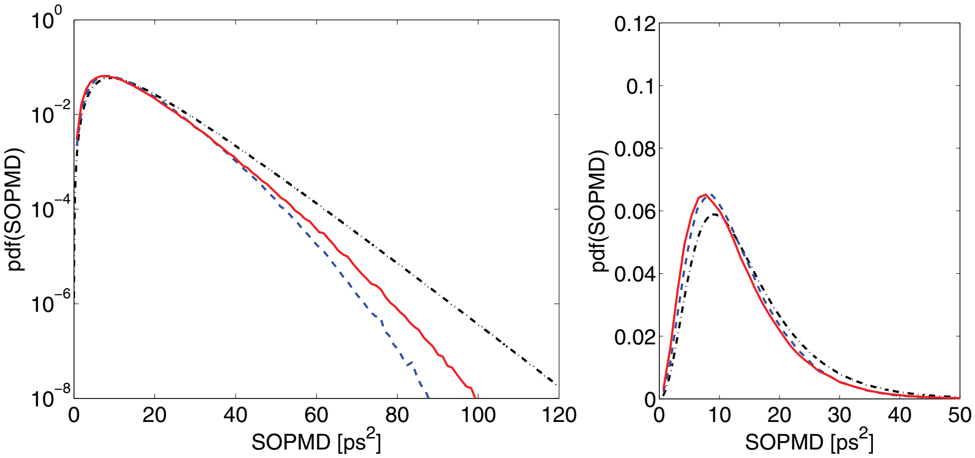


FIG. 4.4. *Same as Figure* 4.3, *but for* $N = 20$ *and with the same section DGDs as in Figure* 4.2.

4.2. PDF of second-order PMD. No analytical solutions exist for the PDF of second-order PMD generated by a concatenation of a finite number of fixed-length sections, even in the isotropic case and even in the case of equal-length sections. Therefore, one must resort to numerical simulations for both the isotropic and the rotator model. We used IS-MC simulations as discussed in section 3.2 with $\Phi_{\max} = \pi/2$, again with a combination of biasing strengths to cover the whole range of second-order PMD. More precisely, we used the same values of biasing strength as when biasing for large values of DGD.

Figure 4.3 shows the PDF of second-order PMD for the isotropic and the rotator model for the same section lengths as in Figure 4.1, while Figure 4.4 does the same for the same section lengths as in Figure 4.2. In all cases the solid line shows the nominal

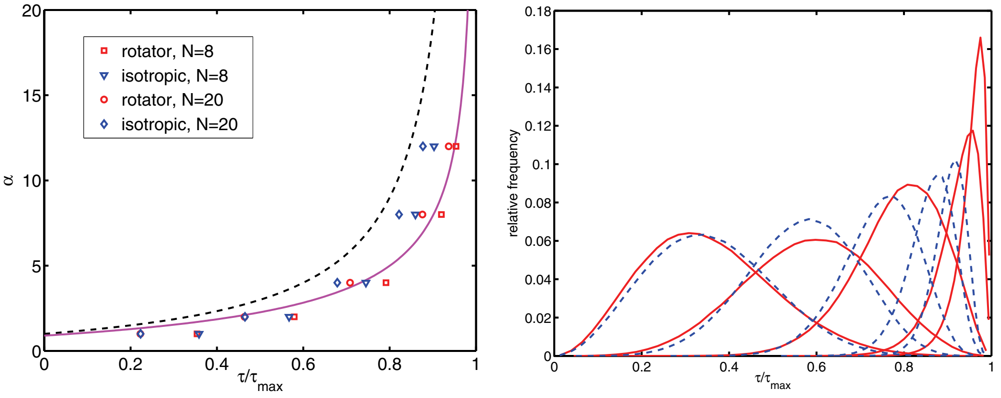


FIG. 4.5. *Left: The estimated value of the biasing strength needed to obtain a given fraction of the maximum DGD. Dashed line: uncorrected value; solid line: value with the heuristic correction. Also shown are the means of the biased MC samples obtained with the biasing strengths* $\alpha = 1, 2, 4, 8, 12$ *for both the isotropic and the rotator model with* $N = 8$ *and* $N = 20$*. Right: The relative frequency of the total DGD obtained with the biasing values for* $N = 8$ *in the isotropic model (dashed curves) and the rotator model (solid curves) in case* A.

"sech-tanh" distribution obtained in the limit of large number of sections [13], namely,

$$p_{\mathrm{sopmd}}(x) = \frac{8}{\pi\langle\tau\rangle^2}\, y(x)\,\mathrm{sech}[y(x)]\,\tanh[y(x)]\,, \tag{4.3}$$

where $y(x) = 4x/\langle\tau\rangle^2$. For both the isotropic and the rotator model, the simulations were done with the same biasing strengths as for the DGD, with 200,000 samples per biasing strength for both models. As with the DGD, the tails of the PDF of the rotator model are significantly larger than those of the isotropic model. (The maximum second-order PMD for an isotropic concatenation of sections was obtained in [18].) Here, however, the difference between the two models seems to be less pronounced than for the PDF of the DGD.

It should be clear from the figures that the overall PDFs of both DGD and second-order PMD can depend significantly on how much of the Poincaré sphere is being sampled by the hinge rotation matrix. It should also be clear that since the maximum DGD and maximum second-order PMD of any PMD emulator are determined by the particular values of the individual DGDs, the resulting PDFs can vary quite a bit for different choices of the section DGDs, even though the individual DGDs are all drawn from an identical Maxwellian distribution. In particular, these PDFs can occasionally differ significantly from the average Maxwellian and sech-tanh distributions, even at moderate values of DGD and second-order PMD, and even with a relatively large number of sections. This is indeed evident from Figures 4.1 and 4.3.

4.3. Targeting. Figure 4.5 shows (to the left) a comparison of the estimate (3.17) of the biasing strength α_o (dashed line) required to obtain a given target DGD versus the heuristic correction α (solid line). Both are plotted against the normalized total DGD (namely, the ratio $\tau/\tau_{\max}$). Also plotted is the mean of the biased MC samples obtained for both the isotropic and the rotator model with $\alpha = 1$ (unbiased), 2, 4, 8, and 12, computed for both $N = 8$ and $N = 20$ with the section lengths listed in Table 4.1. (The results from cases A and B are indistinguishable from each other.) As is evident from this comparison, a good agreement exists between the analytical approximation and the actual mean of the MC samples for both the isotropic and the rotator model for values of N up to 20, except for small values of α. (For $\alpha = 1$, (3.17) yields the unphysical value $\langle\tau\rangle = 0$.) Note, however, that the agreement becomes worse for larger values of N. This is to be expected, since $\langle\tau\rangle/\tau_{\max} \to 0$ in the limit $N \to \infty$, even when the mean DGD per section is scaled as $\langle\tau\rangle/\sqrt{N}$ so as to keep $\langle\tau\rangle$ fixed.

Also shown (to the right) are histograms showing the expected relative frequency of the values of total DGD (as estimated from biased MC simulations) for case A for both the isotropic (dashed lines) and the rotator (solid lines) models for the same values of α as above, illustrating how the IS-MC samples indeed cluster around the expected mean for both the isotropic and the rotator model.

5. Conclusions. We have discussed two models of PMD generation, both consisting of a concatenation of unequal-length sections: a conventional, isotropic model, based on the assumption that the action of the hinges connecting the individual fiber sections causes their relative orientations to vary uniformly across the Poincaré sphere; and a rotator model, based on linearly birefringent elements that rotate relative to one another. We have presented the implementation of IS for both the isotropic model and the rotator model with sections of arbitrary length, and we have used analytical and numerical methods to compute the statistics of first- and second-order PMD in both models. The results show that the PMD statistics depend significantly upon the details of how PMD is generated.

We should reiterate that even though only first- and second-order PMD are biased, a full range of higher-order PMD is also being generated. Moreover, it has been shown [24] that multiple IS with a proper choice of biasing strengths which cover the whole (DGD, second-order PMD) plane is sufficient to accurately capture the statistical distribution of PMD-induced transmission penalties even when multistage PMD compensators are used and even when first-order and second-order PMD are completely compensated.

Thus, the present methods can be employed to compute PMD-induced pulse distortions in systems with various configurations. The present work also provides a further demonstration that the PMD-induced penalties depend on the specific physical details of how PMD is generated in the system. More specifically, for the hinge model they depend on how much of the Poincaré sphere is sampled by the hinge rotation matrix. If the actual PMD generation mechanism in some realistic situation can be considered to be a hybrid between the isotropic and the rotator models, these two models could then provide useful upper and lower limits for the actual penalties in the system.

Acknowledgment. We thank H. Kogelnik and P. J. Winzer for many insightful discussions.

REFERENCES

[1] C. Antonelli and A. Mecozzi, *Statistics of the DGD in PMD emulators*, IEEE Photon. Technol. Lett., 16 (2004), pp. 1804–1806.

[2] C. Antonelli and A. Mecozzi, *Theoretical characterization and system impact of the hinge model of PMD*, J. Lightwave Technol., 24 (2006), pp. 4064–4074.

[3] C. Antonelli, A. Mecozzi, K. Cornick, M. Brodsky, and M. Boroditsky, *PMD-induced penalty statistics in fiber links*, IEEE Photon. Technol. Lett., 17 (2005), pp. 1013–1015.

[4] R. Barakat, *Isotropic random flights*, J. Phys. A, 6 (1973), pp. 796–804.

[5] G. Biondini and W. L. Kath, *PMD emulation with Maxwellian length sections and importance sampling*, IEEE Photon. Technol. Lett., 16 (2004), pp. 789–791.

[6] G. Biondini and W. L. Kath, *Polarization-dependent chromatic dispersion and its impact on return-to-zero transmission formats*, IEEE Photon. Technol. Lett., 17 (2005), pp. 1866–1868.

[7] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization-mode dispersion*, IEEE Photon. Technol. Lett., 14 (2002), pp. 310–312.

[8] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization mode dispersion: Techniques and applications*, IEEE J. Lightwave Technol., 22 (2004), pp. 1201–1215; errata, 24 (2006), p. 1065.

[9] M. Brodsky, M. Boroditsky, P. Magill, N. J. Frigo, and M. Tur, *Persistence of spectral variations in DGD statistics*, Opt. Expr., 13 (2005), pp. 4090–4095.

[10] M. Brodsky, N. J. Frigo, M. Boroditsky, and M. Tur, *Polarization-mode dispersion of installed fibers*, IEEE J. Lightwave Technol., 17 (2006), p. 4584.

[11] S. L. Fogal, G. Biondini, and W. L. Kath, *Multiple importance sampling for first- and second-order polarization mode dispersion*, IEEE Photon. Technol. Lett., 14 (2002), pp. 1273–1275; errata, 14 (2002), p. 1487.

[12] G. J. Foschini and C. D. Poole, *Statistical theory of polarization dispersion in single mode fibers*, IEEE J. Lightwave Technol., 9 (1991), pp. 1439–1456.

[13] G. J. Foschini, L. E. Nelson, R. M. Jopson, and H. Kogelnik, *Probability densities of second-order polarization-mode dispersion including polarization-dependent chromatic dispersion*, IEEE Photon. Technol. Lett., 12 (2000), pp. 293–295.

[14] A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, *Statistical characterization of fiber random birefringence*, Opt. Lett., 25 (2000), pp. 1322–1324.

[15] A. Galtarossa, L. Palmieri, M. Schiano, and T. Tambosso, *Measurements of birefringence correlation length in long single-mode fibers*, Opt. Lett., 26 (2001), p. 962.

[16] J. P. Gordon and H. Kogelnik, *PMD fundamentals: Polarization-mode dispersion in optical fibers*, Proc. Natl. Acad. Sci., 97 (2000), pp. 4541–4550.

[17] M. Karlsson, *Probability density functions of the differential group delay in optical fiber communication systems*, IEEE J. Lightwave Technol., 19 (2001), pp. 324–331.

[18] M. Karlsson, *Geometric interpretation of second-order PMD*, IEEE J. Lightwave Technol., 26 (2006), pp. 643–651.

[19] M. Karlsson, J. Brentel, and P. A. Andrekson, *Long-term measurement of PMD and polarization drift in installed fibers*, J. Lightwave Technol., 18 (2000), pp. 941–951.

[20] H. Kogelnik, L. E. Nelson, and R. M. Jopson, *Polarization mode dispersion*, in Optical Fiber Telecommunications IVB, I. P. Kaminow and T. Li, eds., Academic Press, San Diego, CA, 2002, pp. 725–861.

[21] H. Kogelnik, P. Winzer, L. E. Nelson, R. M. Jopson, M. Boroditsky, and M. Brodsky, *First-order PMD outage for the hinge model*, IEEE Photon. Technol. Lett., 17 (2005), pp. 1208–1210; errata, 17 (2005), p. 2499.

[22] J. H. Lee, M. S. Kim, and Y. C. Chung, *Statistical PMD emulator using variable DGD elements*, IEEE Photon. Technol. Lett., 15 (2003), pp. 54–56.

[23] J. Li, G. Biondini, H. Kogelnik, and P. J. Winzer, *Noncompliant capacity ratio for systems with an arbitrary number of polarization hinges*, IEEE J. Lightwave Technol., 26 (2008), pp. 2110–2117.

[24] A. O. Lima, C. R. Menyuk, and I. T. Lima, *Comparison of two biasing Monte Carlo methods for calculating outage probabilities in systems with multisection PMD compensators*, IEEE Photon. Technol. Lett., 17 (2005), pp. 2580–2582.

[25] I. T. Lima, Jr., A. O. Lima, G. Biondini, C. R. Menyuk, and W. L. Kath, *A comparative study of single-section polarization-mode dispersion compensators*, IEEE J. Lightwave Technol., 22 (2004), pp. 1023–1032.

[26] A. Mecozzi, C. Antonelli, M. Boroditsky, and M. Brodsky, *Characterization of the time dependence of polarization mode dispersion*, Opt. Lett., 29 (2004), pp. 2599–2601.

[27] M. Shtaif, *The Brownian-bridge method for simulating polarization mode dispersion in optical communications systems*, IEEE Photon. Technol. Lett., 15 (2003), pp. 51–53.

[28] Y. Tan, J. Yang, W. L. Kath, and C. R. Menyuk, *Transient evolution of the polarization-dispersion vector's probability distribution*, J. Opt. Soc. Amer. B, 19 (2002), pp. 992–1000.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 565–576

IMPEDANCE-ACOUSTIC TOMOGRAPHY*

BASTIAN GEBAUER† AND OTMAR SCHERZER‡

Abstract. In this work we present a new hybrid imaging technique that combines electrical impedance tomography (EIT) with acoustic tomography. The novel technique makes use of the fact that the absorbed electrical energy inside the body raises its temperature, thus leading to expansion effects. The expansion then induces an acoustic wave which can be recorded outside the body and consequently be used to calculate the absorbed energy inside the body, from which the electrical conductivity can be reconstructed. In other words, we try to combine the high contrast of EIT with the high resolution of ultrasound.

Key words. hybrid imagery, electrical impedance tomography (EIT), thermoacoustics

AMS subject classifications. 47J06, 35R30, 65N21, 92C55

DOI. 10.1137/080715123

1. Introduction.

In electrical impedance tomography (EIT) an electrical voltage $f(x)$ is applied to the surface S of a body B, thus giving rise to an electrical potential $u(x)$ inside the body. One then measures the resulting surface current j and tries to reconstruct the conductivity inside the body from one or several voltage-current pair(s) $(f(x), j(x))$ on the boundary.

The problem is known to be severely ill-posed, and though a number of commercial EIT systems exists, a stable reconstruction algorithm still seems to be out of reach. As a starting point we refer the interested reader to the survey articles of Cheney, Isaacson, and Newell [7], Borcea [4, 5], Lionheart [19], and Bayford [3]. A recent study reporting on the high permittivity and conductivity contrast of different breast tissues over the frequency range 40Hz–100MHz can be found in Stoneman et al. [27]. It has to be noted, however, that the conductivity value alone does not seem to be a sufficient criterion to distinguish cancerous from healthy breast tissue; cf. also the studies of Lazebnik et al. [16, 17] for the microwave frequency range 0.5–20GHz.

A promising approach to overcoming the intrinsic ill-posedness of the problem is to combine EIT with another imaging system that provides additional information. The most prominent example is magnetic resonance electrical impedance tomography (MREIT), which combines EIT with measurements of the magnetic flux from which one obtains the current density *inside* the body; cf. the works of Kwon et al. [14], S. Kim et al. [12], Y. J. Kim et al. [11], and the recent work of Nachman, Tamasan, and Timonov [22]. Another approach is magnetoacoustic imaging, where an exterior magnetic field is used to generate displacements in the body via the Lorentz force. The resulting pressure wave is measured by ultrasound transducers and provides information about the interior current density; cf., e.g., Ma and He [20] and the preprint

*Received by the editors February 5, 2008; accepted for publication (in revised form) August 1, 2008; published electronically December 3, 2008.
http://www.siam.org/journals/siap/69-2/71512.html

†Institut für Mathematik, Johannes Gutenberg-Universität, 55099 Mainz, Germany (gebauer@math.uni-mainz.de). This work was conducted while this author was employed at the Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstr. 69, A-4040 Linz, Austria.

‡Department of Computer Science, University of Innsbruck, Technikerstr. 21a, 6020 Innsbruck, Austria (otmar.scherzer@uibk.ac.at, http://informatik.uibk.ac.at/infmath). The work of this author was supported by the Austrian Science Foundation, project NFN 10505.

[2] of Ammari et al. [1] for recent references. Independently of this work, Ammari et al. have recently proposed using ultrasound to produce localized elastic perturbations to locally change the conductivity inside a body. From the resulting change in the EIT measurements they obtain the energy density inside the body and use this additional information to reconstruct the conductivity.

In this work we propose a new method to obtain additional interior information for EIT by combining it with ultrasound tomography. Similarly to magnetoacoustic imaging, we obtain our additional information from creating a pressure wave inside the body. However, we do not rely on an externally applied magnetic field and the Lorentz force but on thermal expansion. The resulting additional interior information is the same that Ammari et al. obtain by elastic deformations, i.e., the interior energy density.

To be more specific, we will make use of the fact that the absorbed energy inside the body raises its temperature, thus leading to expansion effects. The expansion then induces an acoustic wave which can be recorded outside the body and consequently be used to calculate the absorbed energy inside the body. The advantage of using acoustical rather than electromagnetic measurements for the reconstruction is that we can then choose the excitation frequency small with respect to the speed of electromagnetic waves (so that the model of impedance tomography holds true) but large with respect to the speed of sound (so that the we obtain a high resolution in the reconstructions). In other words, we try to combine the high contrast of EIT with the high resolution of ultrasound. The ideas for this combination stem from thermoacoustic computerized tomography, where a body is illuminated and thus heated up with a short pulse of light and the resulting acoustic pressure wave is recorded. For an introduction to the field of thermoacoustic tomography, we refer the reader to the recent special section in the journal *Inverse Problems* [23]. A survey on experimental setups for thermoacoustic imaging can be found in Xu and Wang [28].

The outline of this work is as follows. We start by developing the model of impedance-acoustic tomography in section 2.1 and study the well-posedness of the resulting direct problem in section 2.2. In section 3 we study the associated inverse problems and derive first reconstruction algorithms. Finally, we show preliminary numerical examples for the simulation of impedance-acoustic tomography as well as for the reconstruction algorithms in section 4.

2. Impedance-Acoustic Computerized Tomography (ImpACT).

2.1. Derivation of the modeling equations.

2.1. Derivation of the modeling equations. If a stationary electrical voltage $f(x)$ is applied to the surface S of a body $B \subset \mathbb{R}^n$, $n = 2$ or $n = 3$, this gives rise to an electrical potential $u(x)$ inside the body. In the state of equilibrium the potential is given by the solution u of

$$\nabla \cdot (\sigma(x) \nabla u(x)) = 0 \qquad \text{in } B, \tag{2.1}$$

$$u(x)|_S = f(x) \qquad \text{on } S, \tag{2.2}$$

where $\sigma(x)$ is the specific conductivity of the body. One can then measure the resulting surface current

$$j(x) = \sigma(x) \partial_\nu u(x)|_S$$

($\nu = \nu(x)$ is the outer normal vector at a surface point $x \in S$) and try to reconstruct $\sigma(x)$ from one or several voltage-current pair(s) $(f(x), j(x))$.

An electrical potential almost instantly reaches its state of equilibrium, so that if we apply a time-dependent voltage $F(x,t) = f(x)\sqrt{g(t)}$ that varies slowly in time (compared to the speed of electromagnetic waves), the induced electric potential is given by its quasi-static limit $U(x,t) = u(x)\sqrt{g(t)}$, where $u(x)$ solves (2.1), (2.2). (We choose the notation $\sqrt{g(t)}$ in order to have g proportional to the amount of applied electrical power in (2.3) below.)

When electrical currents are flowing through a body, three effects can be observed: stimulation of nerves, electrolysis, and thermal heating; cf., e.g, [9]. We concentrate on the heating effect and suggest using high frequency currents, which have less stimulating effects on the nerves. Indeed, the effect of thermal heating with high frequency currents is exploited in *high frequency surgery*; cf., e.g., [9, 13].

We will make use of the usual convention that "∇" and "Δ" are taken only with respect to spatial variables and that first, respectively, second, partial time-derivatives are denoted by one, respectively, two, dots. *Joule's law* describes the relation between the rate of variation of energy Q, i.e., the absorbed electrical power density $\dot{Q}(x,t)$, and the electric potential by

$$\dot{Q}(x,t) = \sigma(x)|\nabla U(x,t)|^2 = \sigma(x)|\nabla u(x)|^2 g(t). \tag{2.3}$$

If the voltage is applied only for a short time, we can neglect thermal diffusion so that the change of temperature $T(x,t)$ is given by

$$\dot{T}(x,t) = \frac{1}{\rho(x,t)c(x)}\dot{Q}(x,t), \tag{2.4}$$

where $c(x)$ is the specific heat capacity (i.e., the amount of energy needed to heat up a unit mass of the material by one unit of temperature) and $\rho(x,t)$ is the mass density.

Before we continue with our modeling equation, let us give a rough quantitative estimate of practically realizable temperature changes. We use the standard SI-units cm, m for centimeters and meters, μs for microseconds, g for grams, Ω for ohm, A for ampere, MHz for megahertz, J for joule, and mK and K, for millikelvin and kelvin. For the specific heat capacity and density we take the values of breast fat from Robinson et al. [24, Table 5], $c = 2.43\text{J}/(\text{gK})$ and $\rho = 0.934\text{g}/\text{cm}^3$. The specific electrical conductivity of adipose tissue at a frequency of 1MHz is about $\sigma = 0.4/(\Omega\text{m})$; cf. [27, Fig. 5]. Thus, a specimen cube of 1cm side length has a mass of $M = 0.934\text{g}$, and its electrical resistance is $R = \sigma^{-1}\frac{\text{length}}{\text{area}} = 250\Omega$. Applying a pulse of $\Delta t = 1\mu\text{s}$ with $\sigma|\nabla u| = I = 3\text{A}$ will change the temperature of this cube by

$$\Delta T = \frac{1}{M}\frac{1}{c}RI^2\Delta t \approx 990\frac{\text{K}\,\Omega\,\text{A}^2\,\mu\text{s}}{\text{J}} = 0.99\,\text{mK}.$$

This temperature rise seems enough to produce ultrasound waves, which can be measured with ultrasound transducers while still being unharmful to biological tissue.

We now proceed as in the derivation of the equations of thermoacoustic tomography in the book of Scherzer et al. [26]; cf. also the publication of Haltmeier, Schuster, and Scherzer [10] or the recent review article of Xu and Wang [28]. The change in the material's temperature can be related to the change in its density ρ and to the change in its pressure p via the so-called (linearized) expansion equation

$$\beta(x)\dot{T}(x,t) = \frac{1}{v_s^2}\dot{p}(x,t) - \dot{\rho}(x,t), \tag{2.5}$$

where v_s is the speed of sound and $\beta(x)$ is the thermal expansion coefficient that specifies the increase of volume per increase of temperature. Under the assumption that the density ρ is only slightly varying from a constant value ρ_0 and only small velocities occur, the velocity v and the density ρ are coupled by the linearized continuity equation

$$\dot{\rho}(x,t) = -\rho_0 \nabla \cdot v(x,t). \tag{2.6}$$

Furthermore, assuming an inviscid, nonturbulent flow of material with just slightly varying pressure, the velocity v is related to the pressure p by the linearized Euler equation

$$\rho_0 \dot{v}(x,t) = -\nabla p(x,t). \tag{2.7}$$

Combining (2.3)–(2.7) and again applying our assumption that the density ρ is only slightly varying from ρ_0, we obtain

$$\frac{1}{v_s^2}\ddot{p}(x,t) - \Delta p(x,t) = \frac{\beta(x)}{\rho_0 c(x)}\sigma(x)|\nabla u(x)|^2 \dot{g}(t).$$

If the electric energy is applied only for a very short time (compared to the speed of sound), we can replace $g(t)$ by a δ-peak and obtain that p is the solution of

$$\begin{aligned} \frac{1}{v_s^2}\ddot{p}(x,t) - \Delta p(x,t) &= 0 && \text{in } \mathbb{R}^n, \\ p(x,0) &= \frac{\beta(x)}{\rho_0 c(x)}\sigma(x)|\nabla u(x)|^2 \chi_B(x) && \text{in } \mathbb{R}^n, \\ \dot{p}(x,0) &= 0 && \text{in } \mathbb{R}^n, \end{aligned}$$

where χ_B is the characteristic function of B. We furthermore assume that the specific heat capacity and the thermal expansion coefficient are approximately constant and known. By a standard change of units we can then eliminate v_s, β, ρ_0, and c from the equations and obtain, together with (2.1) and (2.2), the *equations of impedance-acoustic tomography*:

$$\nabla \cdot \sigma \nabla u(x) = 0 \quad \text{in } B, \tag{2.8}$$

$$u(x)|_S = f \quad \text{on } S, \tag{2.9}$$

$$\ddot{p}(x,t) - \Delta p(x,t) = 0 \quad \text{in } \mathbb{R}^n, \tag{2.10}$$

$$p(x,0) = \sigma(x)|\nabla u(x)|^2 \chi_B(x) \quad \text{in } \mathbb{R}^n, \tag{2.11}$$

$$\dot{p}(x,0) = 0 \quad \text{in } \mathbb{R}^n. \tag{2.12}$$

The *forward problem of impedance-acoustic tomography* can now be stated as follows: Given the conductivity σ and the applied voltage f on S, determine the resulting currents $\sigma\partial_\nu u|_S$ and the resulting pressure wave $p(x,t)$ that solves (2.8)–(2.12). The *inverse problem of impedance-acoustic tomography* consists of reconstructing the conductivity σ from knowledge of the applied voltage f and measurements of the resulting currents $\sigma\partial_\nu u|_S$ and the resulting pressure wave p on some part of $\mathbb{R}^3$. In this work we will restrict ourselves to the case where p is measured on the whole surface S for some time interval $[0,T]$, and we will furthermore assume that σ is known in a small neighborhood of this surface S.

Finishing this subsection, let us recapitulate the crucial assumptions regarding the time-scale in our model and give a rough estimate of feasible parameters. The applied voltage must vary slowly in time compared to the speed of electromagnetic waves, so that the (quasi-static) equations of impedance tomography are valid. The maximum frequency of Rensselaer's ACT 4 EIT system is 1MHz; cf. Saulnier et al. [25]. At the same time, the voltage must be applied only for such a small time that thermal diffusion can be neglected and that the applied energy takes the form of a delta-pulse in the time-scale of sound waves, i.e., also that stress propagation can be neglected during the application of the pulse. The latter two conditions are commonly referred to as *thermal and stress confinements* in thermoacoustic tomography; cf. Xu and Wang [28]. The stress confinement is the more stringent condition and, for a pulse of 1μs, limits the expected spatial resolution to 1.5mm; cf. [28].

2.2. Well-posedness of the direct problem. Throughout this work we will assume that $B \subset \mathbb{R}^n$, $n = 2$ or $n = 3$, is a smoothly bounded domain, $T > 0$, $f \in W^{7/4,4}(S)$, and $\sigma \in W_+^{1,\infty}(B)$, where the subscript $+$ denotes the subspace of functions with positive essential infima. Under this assumption we obtain the following lemma.

LEMMA 2.1. *For every $\sigma \in W_+^{1,\infty}(B)$, there exists a unique solution $u \in W^{2,4}(B)$ of* (2.8) *and* (2.9)*. Setting*

$$\mathcal{E}(\sigma) := \sigma|\nabla u|^2, \qquad \textit{where } u \textit{ solves } (2.8) \textit{ and } (2.9),$$

defines a mapping $\mathcal{E}: W_+^{1,\infty}(B) \to H^1(B)$.

Proof. Note that the space $W^{7/4,4}(S)$ is the space of traces of functions from $W^{2,4}(B)$. From standard regularity results for elliptic equations, we obtain that for every $\sigma \in W_+^{1,\infty}(B)$ there exists a unique solution $u \in W^{2,4}(B)$ of (2.8) and (2.9) (cf., e.g., Miranda [21, Thm. 38,VI]). Thus the result follows from the product rule for Sobolev functions. □

Some caution has to be taken in the treatment of the acoustic equations (2.10)–(2.12). Though we have just seen that our regularity assumptions guarantee that the initial condition is an H^1-function in B, its continuation by zero to $\mathbb{R}^n$ will in general have a jump across S. Roughly speaking, this jump persists in the pressure wave, so that our measurements $p|_S$ are in general not well defined (as a function). Of course, from a practical point of view, there cannot be ambiguity in the measurement data; thus this problem shows that the idealization of a pressure wave appearing instantly in a sharply bounded body is not consistent with the idealization of a measurement surface with zero thickness.

However, using our additional assumption that we know the conductivity σ on S, we can circumvent this problem without giving up one of these two idealizations. The quantity $\sigma(x)|\nabla u(x)|^2|_S$ can be calculated from $\sigma|_S$, the measured surface currents $\sigma(x)\partial_\nu u(x)|_S$, and the applied voltage $u(x)|_S = f(x)$. Using a function $\tilde{p}_0(x) \in H^1(B)$ with the same boundary values $\tilde{p}_0(x)|_S = \sigma(x)|\nabla u(x)|^2|_S$, we can define the solution $\tilde{p}$ of

$$\ddot{\tilde{p}}(x,t) - \Delta\tilde{p}(x,t) = 0 \qquad \text{in } \mathbb{R}^n, \tag{2.13}$$

$$\tilde{p}(x,0) = \tilde{p}_0(x)\chi_B(x) \qquad \text{in } \mathbb{R}^n, \tag{2.14}$$

$$\dot{\tilde{p}}(x,0) = 0 \qquad \text{in } \mathbb{R}^n. \tag{2.15}$$

The difference $q := p - \tilde{p}$ then solves the wave equation with an initial condition in $W^1(\mathbb{R}^n)$, and thus its trace on S is well defined. Before we state this in a rigorous form below, let us comment on the practical realization of this approach. One can easily compute a smooth approximation to $\tilde{p}$ and evaluate this on S. The difference of the measurement of p on S and this quantity can then be regarded as an approximation to the well-defined, idealized model measurements $q|_S$.

We now restate the above arguments in a rigorous form.

Lemma 2.2. *Denote by*

$$\gamma:\ H^1(B) \to H^{1/2}(S), \qquad v \mapsto v|_S,$$

the trace operator on S*, and let* $\gamma^-:\ H^{1/2}(S) \to H^1(B)$ *be a continuous right inverse of* γ*, i.e.,* $\gamma\gamma^- = Id$*.*

For every $\sigma \in W^{1,\infty}_+(B)$ *there exists a unique solution* $\tilde{p} \in C(0,T,L^2(\mathbb{R}^n))$ *of* (2.13)–(2.15) *with* $\tilde{p}_0 := \gamma^-\gamma\mathcal{E}(\sigma)$ *and a unique solution* $p \in C(0,T,L^2(\mathbb{R}^n))$ *of* (2.10) *and* (2.12) *with* $p(x,0) = \mathcal{E}(\sigma)\chi_B$*. Their difference* $q := p - \tilde{p}$ *is an element of* $C(0,T,W^1(\mathbb{R}^n))$ *and it is the unique solution of*

$$\ddot{q}(x,t) - \Delta q(x,t) = 0 \qquad \text{in } \mathbb{R}^n, \tag{2.16}$$

$$q(x,0) = q_0(x)\chi_B(x) \qquad \text{in } \mathbb{R}^n, \tag{2.17}$$

$$\dot{q}(x,0) = 0 \qquad \text{in } \mathbb{R}^n, \tag{2.18}$$

with $q_0 = (I - \gamma^-\gamma)\mathcal{E}(\sigma) \in H^1_0(B)$*. Also, the mapping*

$$\mathcal{F}:\ H^1_0(B) \to C(0,T,H^{1/2}(S)), \qquad q_0 \mapsto q|_S, \qquad \textit{where } q \textit{ solves } (2.16)\text{–}(2.18),$$

is continuous and linear.

Proof. This follows from classical results on the wave equation; cf., e.g., Lions and Magenes [18, Chp. 3, Thm. 9.3] for the unique existence of $\tilde{p}, p \in C(0,T,L^2(\mathbb{R}^n))$ and [18, Chp. 3, Thm. 8.2] for the unique existence of a solution $q \in C(0,T,W^1(\mathbb{R}^n))$ of (2.16)–(2.18). □

3. Inverse problems of ImpACT.

In the last section we saw that, using the known boundary values of the conductivity $\sigma|_S$, the measured surface currents $\sigma(x)\partial_\nu u(x)|_S$, and the applied voltage $f(x)$, we can calculate $\gamma\mathcal{E}(\sigma) = \sigma|\nabla u|^2|_S$ and thus the *modified pressure measurements* $q|_S$ from the real measurement data. The dependence of $q|_S$ from the unknown conductivity σ is given by

$$q|_S = \mathcal{F}(I - \gamma^-\gamma)\mathcal{E}(\sigma).$$

Hence, the inverse problems of determining the conductivity from the measurements leads to the problems of inverting the two linear operators $\mathcal{F}$, $(I - \gamma^-\gamma)$ and the nonlinear operator $\mathcal{E}$. Since $\gamma\mathcal{E}(\sigma)$ is known, the inversion of $(I - \gamma^-\gamma)$ consists simply of adding $\gamma^-\gamma\mathcal{E}(\sigma)$. It therefore remains to invert $\mathcal{F}$, i.e., to determine the initial value of a pressure wave from its trace on S, and to invert $\mathcal{E}$, i.e., to determine the conductivity of a body from its electrical energy density.

3.1. Determining the initial condition of the pressure wave.

Recall that $\mathcal{F}:\ H^1_0(B) \to C(0,T,H^{1/2}(S))$ maps the initial value $q_0 \in H^1_0$ to the boundary values $q|_S$, where $q \in C(0,T,H^1(\mathbb{R}^n))$ solves

$$\begin{aligned} \ddot{q}(x,t) - \Delta q(x,t) &= 0 && \text{in } \mathbb{R}^n,\\ q(x,0) &= q_0(x)\chi_B(x) && \text{in } \mathbb{R}^n,\\ \dot{q}(x,0) &= 0 && \text{in } \mathbb{R}^n. \end{aligned}$$

The inverse problem of determining q_0 from $q|_S$ has been studied to some extent in the context of thermoacoustical tomography. We will use a conceptionally simple time-reversal algorithm that is also mentioned in the work of Finch, Patch, and Rakesh [8, Thm. 5] and for which Burgholzer et al. show numerical results in the recent work [6]. Assume that $T > \operatorname{diam}(B)$; then in the case of three spatial dimensions, the pressure wave q will have completely left the body B, so that $q(x,T)|_B = 0$ and $\dot{q}(x,T)|_B = 0$. Thus the time-reversed pressure wave $r(x,t) := q(x,T-t)|_B$ solves the initial boundary value problem

$$\ddot{r}(x,t) - \Delta r(x,t) = 0 \quad \text{in } B, \tag{3.1}$$

$$r(x,t)|_S = g \quad \text{on } S, \tag{3.2}$$

$$r(x,0) = 0 \quad \text{in } B, \tag{3.3}$$

$$\dot{r}(x,0) = 0 \quad \text{in } B, \tag{3.4}$$

where the boundary data $g = q(x,T-t)|_S$ are just the time-reversed measurements.

Lemma 3.1. *For every $g \in C(0,T,H^{1/2}(S))$, there exists a unique solution $r \in C(0,T,L^2(B))$ to* (3.1)–(3.4).

Proof. Under the weaker assumption that $g \in L^2(0,T,L^2(S))$, this was shown by Lasiecka, Lions, and Triggiani in [15, Thm. 2.3]. □

Since $q(x,T-t)$ solves (3.1)–(3.4), the initial condition $q(x,0)|_B = r(x,T)$ can thus be reconstructed from $q(x,t)|_S$ by solving (3.1)–(3.4) with $g = q(x,T-t)|_S$. Physically this can be interpreted as a combination of a time-reversal of waves and the restriction to a bounded domain.

In the case of two spatial dimensions, the wave does not leave the body completely; thus the solution of (3.1)–(3.4) will not completely agree with $q(x,T-t)$. However, if T is chosen large enough, then only a small part of the wave will still be in B, so that one can expect that the above method will still yield a good approximation to $q(x,0)$.

3.2. Determining the conductivity from the electrical energy. We now turn to the determination of the conductivity from the electrical energy, i.e., to the inversion of the nonlinear mapping

$$\mathcal{E}: \ W_+^{1,\infty}(B) \to H^1(B), \qquad \sigma \mapsto \sigma(x)|\nabla u(x)|^2,$$

where u solves

$$\nabla \cdot \sigma \nabla u(x) = 0 \quad \text{in } B, \quad \text{and} \quad u|_S = f.$$

(Note we keep the applied voltage f fixed throughout this work.)

In [1], Ammari et al. reformulate this problem using the 0-Laplacian and propose an iterative reconstruction strategy that relies on the measurement of two different current patterns. We derive here a similar iterative scheme that is based on a formal Newton algorithm. Denote by u_σ the solution of

$$\nabla \cdot \sigma \nabla u = 0, \qquad u|_S = f.$$

It is well known (and easily shown) that (for $\operatorname{supp} \tau \subset B$) the directional derivative

$$v_\tau := \lim_{h \to 0} \frac{u_{\sigma + h\tau} - u_\sigma}{h}$$

is the solution of

$$\nabla \cdot \sigma \nabla v_\tau = -\nabla \cdot \tau \nabla u_\sigma, \qquad v|_S = 0. \tag{3.5}$$

It follows immediately that

$$\mathcal{E}'(\sigma)\tau = \tau |\nabla u_\sigma|^2 + 2\sigma \nabla u_\sigma \cdot \nabla v_\tau.$$

If $\hat{E} = \hat{\sigma}(x)|\nabla u_{\hat{\sigma}}(x)|^2$ is the reconstructed energy density and σ_n is an approximation to the true conductivity $\hat{\sigma}$, then a Newton-step would consist of solving

$$\mathcal{E}'(\sigma_n)\delta = \hat{E} - \sigma_n |\nabla u_{\sigma_n}(x)|^2$$

and the update $\sigma_{n+1} = \sigma_n + \delta$.

To get around the computationally expensive inversion of $\mathcal{E}'(\sigma)$, we can split it into two parts,

$$\mathcal{E}'(\sigma)\tau = (M_\sigma + P_\sigma)\tau,$$

with

$$M_\sigma \tau := \tau |\nabla u_\sigma|^2 \quad \text{and} \quad P_\sigma \tau := 2\sigma \nabla u_\sigma \cdot \nabla v_\tau.$$

Apart from the problem that $|\nabla u|$ might be zero, the inversion of the multiplication operator M is computationally easy. Instead of using the exact inverse $(M_\sigma + P_\sigma)^{-1}$, one can use the approximate inverse $(I - M_\sigma^{-1} P_\sigma) M_\sigma^{-1}$, which is justified when $M_\sigma^{-1} P_\sigma$ is small. Notably, this approximation results in the same algorithm that Ammari et al. propose in [1] and that is therein motivated by a 0-Laplacian formulation:

Given $\hat{E}$, f, and σ_n,

- calculate ∇u_{σ_n},
- set $\tau := \frac{\hat{E}}{|\nabla u_\sigma|^2} - \sigma_n$,
- calculate the solution v_τ of (3.5),
- update $\sigma_{n+1} := \frac{\hat{E} - 2\sigma \nabla u_\sigma \cdot \nabla v_\tau}{|\nabla u_\sigma|^2}$.

4. Numerical examples. We have tested our inversion algorithm on simulated two-dimensional data. The left side of Figure 4.1 shows the exact conductivity distribution σ that we chose as our test example. A background conductivity of 1 is distorted by two discs centered at $(-0.4, -0.15)$ and $(0.4, 0.15)$ in which the conductivity is given by

$$1 + A_j \exp(R_j^{-1} - (R_j^2 - \rho_j^2)^{-1/2}), \quad j = 1, 2,$$

where $R_1 = R_2 = 0.3$ are the radii of the two discs, ρ_j is the respective distance to the center of the jth disc, and $A_1 = 2$, $A_2 = 0.5$, so that the conductivity is smoothly raised to 3, respectively, lowered to 0.5, inside the discs.

Using the commercial finite element software Comsol, we calculated the corresponding electrical energy $\mathcal{E}(\sigma)$ and evaluated it using linear interpolation on the part of an equidistant 200×200 grid on $[-1, 1]^2$ that belongs to the unit circle. $\mathcal{E}(\sigma)$ is shown on the right side of Figure 4.1.

As the continuous right inverse γ^- of the trace operator we take the solution operator for the Dirichlet operator for the Laplace equation, which is implemented by expanding $\gamma\mathcal{E}(\sigma)$ into the L^2-orthonormal functions

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \sin(m\varphi), \frac{1}{\sqrt{\pi}} \cos(m\varphi) \;\middle|\; m = 1, \ldots, 100 \right\}$$

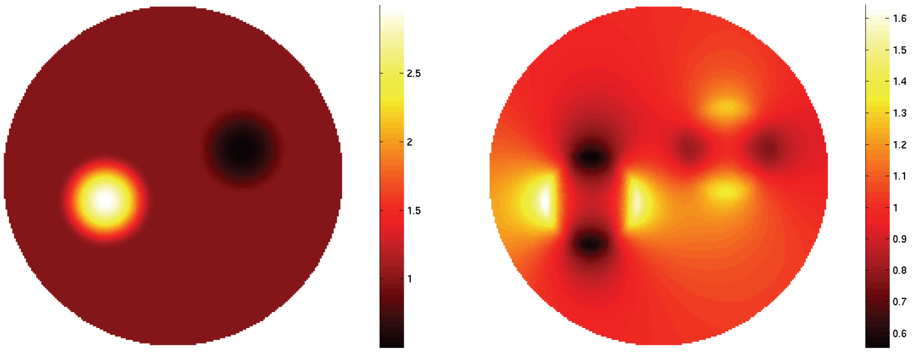


FIG. 4.1. *Exact conductivity and electrical energy distribution.*

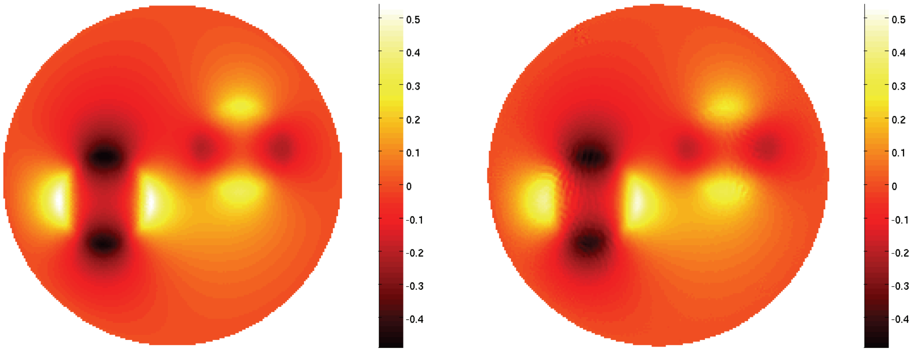


FIG. 4.2. *Exact and reconstructed (modified) energy distribution.*

and using the analytical solutions for the corresponding Dirichlet problems

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \sin(m\varphi) r^m, \frac{1}{\sqrt{\pi}} \cos(m\varphi) r^m \;\middle|\; m = 1, \dots, 100 \right\},$$

where (r, φ) denote the polar coordinates. Accordingly, the left side of Figure 4.2 shows the quantity $(I - \gamma^- \gamma)\mathcal{E}(\sigma)$.

The operator $\mathcal{F}$ is simulated by solving the wave equation with a standard central finite difference scheme on a sufficiently large domain. The values of q on the boundary S are then evaluated using linear interpolation on an equidistant grid on S. Thus, simulated (modified) measurements $q|_S$ are obtained.

We then tested our inversion algorithm on these simulated measurements. The Dirichlet problem for the wave equation was solved using the commercial finite element software Comsol. The reconstructed distribution $r(x, T)$, $T = 4$, is then evaluated using linear interpolation on the part of an equidistant 200×200 grid on $[-1, 1]^2$ that belongs to the unit circle. Figure 4.2 compares the exact (modified) energy distribution $q(x, 0) = (I - \gamma^- \gamma)\mathcal{E}(\sigma)$ (on the left side) with the reconstructed distribution $r(x, T)$, $T = 4$ (on the right side). Figure 4.3 shows profiles of the exact (solid black line) and of the reconstructed energy (dashed red line) on the x-axis (left plot) and on a line connecting the peaks of the true conductivity.

The known quantity $\gamma^- \gamma \mathcal{E}(\sigma)$ is then added to $r(x, T)$, and the Newton algorithm described in section 3.2 is used on this data. As an initial guess we used a constant

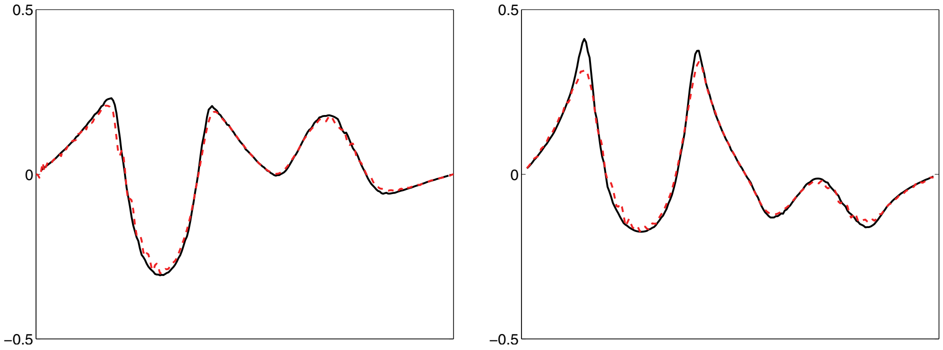


FIG. 4.3. *Profiles of exact and reconstructed (modified) energy.*

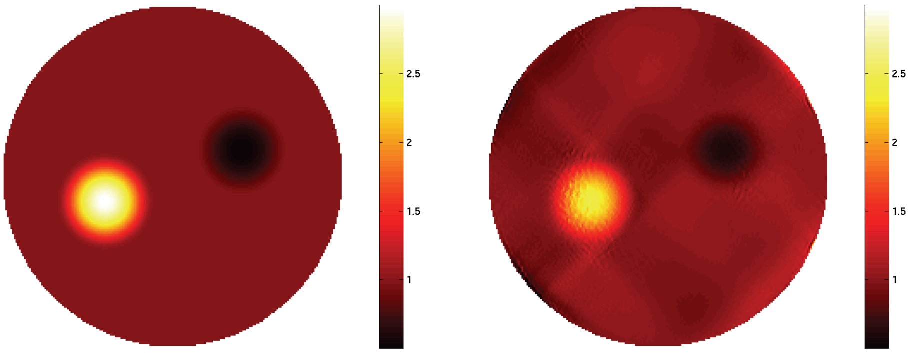


FIG. 4.4. *Exact and reconstructed conductivity distribution.*

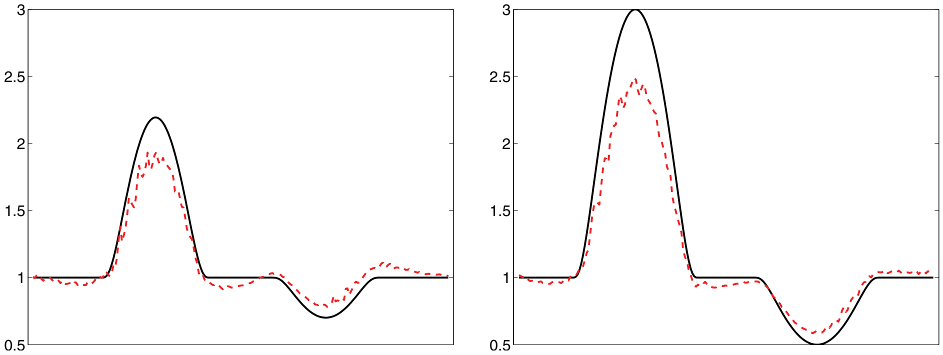


FIG. 4.5. *Profiles of exact and reconstructed conductivity.*

conductivity of 1. The equations appearing in the Newton algorithm were again solved using the commercial finite element software Comsol. Note that thus the same finite element grid is used for the simulation of $\mathcal{E}$ as well as for its inversion. However, the energy is not given, respectively, evaluated, on this grid but on the independent equidistant grid described above, which minimizes the risk of an inverse crime. For the convenience of the reader we show on the left side of Figure 4.4 again the true conductivity to compare it with the best reconstruction that was obtained in the 24th

Newton step (shown on the right side). We observed that the reconstructions do not improve afterward, which seems to be due to accumulated errors. Figure 4.5, which is organized in the same way as Figure 4.3, compares profiles of the true conductivity with the reconstruction.

Acknowledgments. We thank Prof. Dr. Günther Paltauf from the University of Graz for advising us on the practical feasibility of our method and for providing us with information about high frequency surgery. Otmar Scherzer also thanks Richard Kowar from the Infmath Imaging Group of the University of Innsbruck for support with physical quantities.

REFERENCES

[1] H. Ammari, E. Bonnetier, Y. Capdeboscq, M. Tanter, and M. Fink, *Electrical impedance tomography by elastic deformation*, SIAM J. Appl. Math., 68 (2008), pp. 1557–1573.

[2] H. Ammari, Y. Capdeboscq, H. Kang, and A. Kozhemyak *Mathematical models and reconstruction methods in magneto-acoustic imaging*, preprint, available online at http://hal.inria.fr/docs/00/28/85/29/PDF/ACKK_MAT_07_final.pdf.

[3] R. H. Bayford, *Bioimpedance tomography (electrical impedance tomography)*, Ann. Rev. Biomed. Engrg., 8 (2006), pp. 63–91.

[4] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. R99–R136.

[5] L. Borcea, *Addendum to "Electrical impedance tomography,"* Inverse Problems, 19 (2003), pp. 997–998.

[6] P. Burgholzer, M. Haltmeier, G. J. Matt, and G. Paltauf, *Exact and approximative imaging methods for photoacoustic tomography using an arbitrary detection surface*, Phys. Rev. E, 75 (2007), 046706-1.

[7] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[8] D. Finch, S. K. Patch, and Rakesh, *Determining a function from its mean values over a family of spheres*, SIAM J. Math. Anal., 35 (2004), pp. 1213–1240.

[9] R. Haag, *Hochfrequenzchirurgie*, in Medizintechnik: Verfahren - Systeme - Informationsverarbeitung, 2nd ed., Springer-Verlag, Berlin, 2002, pp. 396–411.

[10] M. Haltmeier, T. Schuster, and O. Scherzer, *Filtered backprojection for thermoacoustic computed tomography in spherical geometry*, Math. Methods Appl. Sci., 28 (2005), pp. 1919–1937.

[11] Y. J. Kim, O. Kwon, J. K. Seo, and E. J. Woo, *Uniqueness and convergence of conductivity image reconstruction in magnetic resonance electrical impedance tomography*, Inverse Problems, 19 (2003), pp. 1213–1225.

[12] S. Kim, O. Kwon, J. K. Seo, and J.-R. Yoon, *On a nonlinear partial differential equation arising in magnetic resonance electrical impedance tomography*, SIAM J. Math. Anal., 34 (2002), pp. 511–526.

[13] R. Kramme, *Medizintechnik: Verfahren - Systeme - Informationsverarbeitung*, 2nd ed., Springer-Verlag, Berlin, 2002.

[14] O. Kwon, E. J. Woo, J.-R. Yoon, and J. K. Seo, *Magnetic resonance electrical impedance tomography (MREIT): Simulation study of J-substitution algorithm*, IEEE Trans. Biomed. Engrg., 49 (2002), pp. 160–167.

[15] I. Lasiecka, J.-L. Lions, and R. Triggiani, *Non homogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149–192.

[16] M. Lazebnik, L. McCartney, D. Popovic, C. B. Watkins, M. J. Lindstrom, J. Harter, S. Sewall, A. Magliocco, J. H. Booske, M. Okoniewski, and S. C. Hagness, *A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries*, Phys. Med. Biol., 52 (2007), pp. 2637–2656.

[17] M. Lazebnik, D. Popovic, L. McCartney, C. B. Watkins, M. J. Lindstrom, J. Harter, S. Sewall, T. Ogilvie, A. Magliocco, T. M. Breslin, W. Temple, D. Mew, J. H. Booske, M. Okoniewski, and S. C. Hagness, *A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries*, Phys. Med. Biol., 52 (2007), pp. 6093–6115.

[18] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* I, Grundlehren Math. Wiss. 181, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[19] W. R. B. Lionheart, *EIT reconstruction algorithms: Pitfalls, challenges and recent developments*, Physiol. Meas., 25 (2004), pp. 125–142.

[20] Q. Ma and B. He, *Investigation on magnetoacoustic signal generation with magnetic induction and its application to electrical conductivity reconstruction*, Phys. Med. Biol., 52 (2007), pp. 5085–5099.

[21] C. Miranda, *Partial Differential Equations of Elliptic Type*, Ergeb. Math. Grenzgeb. 2, Springer-Verlag, Berlin, 1970.

[22] A. Nachman, A. Tamasan, and A. Timonov, *Conductivity imaging with a single measurement of boundary and interior data*, Inverse Problems, 23 (2007), pp. 2551–2563.

[23] S. K. Patch and O. Scherzer, eds., *Special section on photo- and thermoacoustic imaging*, Inverse Problems, 23 (2007), pp. S1–S122.

[24] M. P. Robinson, M. J. Richardson, J. L. Green, and A. W. Preece, *New materials for dielectric simulation of tissues*, Phys. Med. Biol. 36 (1991), pp. 1565–1571.

[25] G. J. Saulnier, N. Liu, C. Tamma, H. Xia, T. Kao, J. C. Newell, and D. Isaacson, *An electrical impedance spectroscopy system for breast cancer detection*, in Engineering in Medicine and Biology Society 2007, EMBS 2007: 29th Annual International Conference of the IEEE, IEEE, Washington, DC, 2007, pp. 4154–4157.

[26] O. Scherzer, M. Grasmaier, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging*, Springer-Verlag, New York, in press.

[27] M. R. Stoneman, M. Kosempa, W. D. Gregory, C. W. Gregory, J. J. Marx, W. Mikkelson, J. Tjoe, and V. Raicu, *Correction of electrode polarization contributions to the dielectric properties of normal and cancerous breast tissues at audio/radiofrequencies*, Phys. Med. Biol., 52 (2007), pp. 6589–6604.

[28] M. Xu and L. Wang, *Photoacoustic imaging in biomedicine*, Rev. Sci. Instrum., 77 (2006), 041101.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 577–595

INCLUSION PAIRS SATISFYING ESHELBY'S UNIFORMITY PROPERTY[*]

HYEONBAE KANG[†], EUNJOO KIM[‡], AND GRAEME W. MILTON[§]

Abstract. Eshelby conjectured that if for a given uniform loading the field inside an inclusion is uniform, then the inclusion must be an ellipse or an ellipsoid. This conjecture has been proved to be true in two and three dimensions provided that the inclusion is simply connected. In this paper we provide an alternative proof of Cherepanov's result that an inclusion with two components can be constructed inside which the field is uniform for any given uniform loading for two-dimensional conductivity or for antiplane elasticity. For planar elasticity, we show that the field inside the inclusion pair is uniform for certain loadings and *not* for others. We also show that the polarization tensor associated with the inclusion pair lies on the lower Hashin–Shtrikman bound, and hence the conjecture of Pólya and Szegö is not true among nonsimply connected inclusions. As a consequence, we construct a simply connected inclusion, which is nothing close to an ellipse, but in which the field is almost uniform.

Key words. Eshelby's conjecture, Pólya–Szegö conjecture, uniformity property, inclusions with multiple components, polarization tensor, Weierstrass zeta function

AMS subject classification. 74M25

DOI. 10.1137/070691358

1. Introduction.

Consider a conducting or elastic inclusion subject to a uniform applied field. For certain shapes of inclusions the field inside the inclusion is also uniform, and if this is the case, we say the inclusion has *Eshelby's uniformity property.* Eshelby showed in [9] that ellipses and ellipsoids have the uniformity property and conjectured in [10] that these are the only inclusions with the uniformity property. See also [15]. This conjecture of Eshelby has been proved to be true within the class of simply connected inclusions by Sendeckyj [28] for planar elasticity and by Ru and Schiavone [26] for antiplane elasticity or, equivalently, for two-dimensional conductivity. Recently, a completely different proof of the Eshelby conjecture in two dimensions based on the hodographic transformation was given by Kang and Milton [18]. In the same paper, Eshelby's conjecture in three dimensions was resolved as well. They showed that if a simply connected inclusion with Lipschitz boundary has the uniformity property, then the inclusion must take the shape of an ellipse or an ellipsoid. Independently, Liu (private communication) also established this. As a consequence of Eshelby's conjecture, the conjecture of Pólya and Szegö [25], which asserts that the domain whose polarization tensor has the minimal trace is a disk or a ball, is also proved [17].

Finding a structure inside which the field is uniform is important in the study of composite materials since such a property is required in order to reduce the internal stress of the structure [31]. In fact, it was proved by Grabovsky and Kohn [12] that

[*] Received by the editors May 14, 2007; accepted for publication (in revised form) August 4, 2008; published electronically December 3, 2008. The research of the first two authors was partly supported by grant KOSEF R01-2006-000-10002-0.

http://www.siam.org/journals/siap/69-2/69135.html

[†] Department of Mathematics, Inha University, Incheon 402-751, Korea (hbkang@inha.ac.kr).

[‡] Department of Mathematics, Ewha Womans University, Seoul 120-750, Korea (kej@ewha.ac.kr).

[§] Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (milton@math.utah.edu). This author's research was supported by the National Science Foundation through grant DMS-0411035.

ellipses are the low volume fraction limit of the periodic Vigdergauz microstructure [29, 30], which contains a single inclusion per unit cell. The Vigdergauz microstructure is known to have minimal internal stress among periodic composites. There are also periodic geometries, based on the construction of Hashin [14], that contain a countable number of disks in the unit cell, having Eshelby's uniformity property as follows from section 4 of [6].

In this paper we continue our investigation on the shape of inclusions with the uniformity property. The primary concern of this paper is the construction of inclusions (structures) with two components having smooth boundaries which satisfy Eshelby's uniformity property. This was first solved by Cherepanov [8], and here we provide an alternative proof of Cherepanov's results and give explicit numerical computations of the inclusion shapes.

Another closely related question considered here is whether Eshelby's conjecture is true in a "practical" sense: If the field inside the inclusion is very close to being uniform in some sense, does it follow that the inclusion is very close to an ellipse? (By close to an ellipse we specifically mean that the symmetric difference of the inclusion and an ellipse has small measure.) It is a question of stability.

We construct, in a mathematically rigorous way, inclusions with two components inside which the field is uniform. Figure 2.1 in section 2 shows typical shapes of the inclusion pair. The field inside the inclusion is uniform for any uniform loading in the case of antiplane elasticity, as will be proved in section 3. In the case of linear elasticity, the field is uniform for certain loadings and *not* for other loadings. Using these inclusions, we also answer the question of stability. If we connect two components of the inclusion by a thin bridge, as in Figure 5.1, the field does not change much while the bridged inclusion is simply connected, but far from the shape of an ellipse. In order to construct the structures in this paper, we use the Weierstrass zeta function and the Schwarz–Christoffel formula to solve the free boundary problem. The method of construction in this paper is similar to that of Vigdergauz [29, 30] and Grabovsky and Kohn [12], where the Weierstrass $\mathcal{P}$-function is used to construct the Vigdergauz microstructure.

Eshelby's uniformity property is closely related to the conjecture of Pólya and Szegö on the polarization tensor. In [17] Kang and Milton showed that the polarization tensor satisfies the lower Hashin–Shtrikman bound; then the field inside the inclusion must be uniform, and thus the inclusion is an ellipse provided that it is simply connected. See section 4 for the Hashin–Shtrikman bounds on the polarization tensor. The Pólya–Szegö conjecture follows as an immediate consequence of it. It turns out that the polarization tensor associated with the structure constructed in this paper satisfies the lower Hashin–Shtrikman bound. Therefore, the Pólya–Szegö conjecture does not hold among nonsimply connected inclusions. In the same way as above we are also able to show that stability for the Pólya–Szegö conjecture fails to hold among simply connected inclusions: the bridged inclusion is nothing close to a disk, but the trace of its polarization tensor is very close to being minimal.

This paper is organized as follows: In section 2, we construct inclusions with two components using the Weierstrass zeta function and the Schwarz–Christoffel formula. In section 3, we show that these inclusions enjoy the uniformity property for antiplane elasticity. Section 4 shows that the polarization tensor of the inclusions satisfies the lower Hashin–Shtrikman bound, and hence the Pólya–Szegö conjecture fails to be true among nonsimply connected inclusions. In section 5, we discuss the instability of the uniformity property by connecting the inclusion pair by a thin bridge. In section 6, we analyze the planar elasticity case. We prove that the field inside the inclusion is uniform for certain types of loadings and then show, by numerical computations, that the field is not uniform for some other types of loadings.

2. Construction of the inclusions.

This paper is concerned with a structure consisting of two components, each with a smooth (specifically, Lipschitz) boundary, which satisfy Eshelby's uniformity property in antiplane elasticity or for two-dimensional conductivity. More precisely, we construct an inclusion with two components, B_1 and B_2, such that the solution u to the problem

$$\begin{cases} \nabla \cdot \big(1+(k-1)\chi(B_1\cup B_2)\big)\nabla u = 0 \quad \text{in } \mathbb{R}^2, \\ u(x,y) - a\cdot(x,y) = O(r^{-1}) \quad \text{as } r\to\infty \end{cases} \tag{2.1}$$

is such that ∇u is constant in $B_1\cup B_2$. Here $\chi(B_1\cup B_2)$ is the indicator function of $B_1\cup B_2$, a is a constant vector representing the direction of the uniform loading, and $r=\sqrt{x^2+y^2}$. The conductivity coefficient $1+(k-1)\chi(B_1\cup B_2)$ in (2.1) indicates that the conductivity of the inclusion $B_1\cup B_2$ is $k\neq 1$ while that of the background $\mathbb{R}^2\setminus\overline{B_1\cup B_2}$ is 1. It is worth mentioning that since ∇u is constant (and not 0) in $B_1\cup B_2$, ∂B_1 and ∂B_2 are analytic due to a regularity result of Alessandrini and Isakov [2, Corollary 2.2].

In order to construct such inclusions B_1 and B_2, we will construct a holomorphic function f in $\mathbb{C}\setminus\overline{B_1\cup B_2}$ satisfying

$$f(z) = \Re(cz) + q_j, \quad z\in\partial B_j, \tag{2.2}$$

for some complex constants c and q_j, $j=1,2$, and

$$f(z) = \alpha z + O(1) \quad \text{as } |z|\to\infty \tag{2.3}$$

for some complex number α. Here and afterward, we identify z with $x+iy$. Let us first briefly see why it is enough to construct such a function.

Suppose that there are such simply connected inclusions B_1 and B_2, and let u be the solution to (2.1). Let $u^e := u|_{\mathbb{R}^2\setminus\overline{B_1\cup B_2}}$ and $u^i := u|_{B_1\cup B_2}$. Then there exist holomorphic functions U^e in $\mathbb{C}\setminus\overline{B_1\cup B_2}$ and U^i in $B_1\cup B_2$ such that $\Re U^e = u^e$ and $\Re U^i = u^i$. To see the existence of U^e, it suffices to note that $\int_C \frac{\partial u^e}{\partial\nu}ds = 0$ for any closed piecewise C^1-curve C in $\mathbb{C}\setminus\overline{B_1\cup B_2}$, which can be easily verified using Green's theorem. By (2.1), the solution u satisfies the transmission conditions along the interface ∂B_1 and ∂B_2:

$$u|_+ = u|_- \quad \text{and} \quad \frac{\partial u}{\partial\nu}\Big|_+ = k\frac{\partial u}{\partial\nu}\Big|_- \quad \text{on } \partial B_j,\ j=1,2, \tag{2.4}$$

where the subscripts + and − denote the limits from outside and inside ∂B_j, respectively. It then follows from the Cauchy–Riemann equation that

$$\frac{k+1}{2}U^i - \frac{k-1}{2}\overline{U^i} = U^e + i\lambda_j \quad \text{on } \partial B_j,\ j=1,2, \tag{2.5}$$

for some real constant λ_j. See [16]. Since u^i is linear in each B_j, so is U^i, say,

$$U^i(z) = b_j z + d_j, \quad z\in B_j,\ j=1,2.$$

The constancy of ∇u in $B_1 \cup B_2$ implies $b_1 = b_2 (= b)$. Then (2.5) takes the form

$$\frac{k+1}{2}(bz + d_j) - \frac{k-1}{2}(\overline{bz} + \overline{d}_j) = U^e(z) + i\lambda_j \quad \text{on } \partial B_j, \ j = 1, 2. \tag{2.6}$$

If we put

$$f(z) = U^e(z) - kbz, \tag{2.7}$$

then f is holomorphic in $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ and satisfies (2.2) and (2.3). By reversing the previous arguments we see that if B_1 and B_2 admit a holomorphic function satisfying (2.2) and (2.3), then $B_1 \cup B_2$ has the uniformity property.

For the rest of this section we deal with the problem of constructing two inclusions B_1 and B_2 which admit a function f holomorphic in $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ satisfying (2.2) and (2.3). It turns out that this problem was solved by Cherepanov [8] in a more general setting. Cherepanov showed that there are inclusions with an arbitrary number of components which admit a holomorphic function (outside the inclusion) satisfying (2.2) and (2.3), and then constructed such inclusions with single and double components. The construction of this paper is different from that of [8], and it is more elementary using the explicit formula of the Weierstrass zeta function and the Schwarz–Christoffel formula.

Suppose that f is a holomorphic function in $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ satisfying (2.2) and (2.3). Since such an f maps $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ onto the complex plane with two slits, it is natural to construct an appropriate holomorphic function G on the complex plane with two slits and then define f as the hodographic transform (or the inverse) of G. The use of hodographic transforms is a well-known technique for solving free boundary problems.

Let $0 < a < b$ be two fixed real constants and consider the complex plane with two slits $[-b, -a]$ and $[a, b]$. We first construct a holomorphic function F so that its real parts are constant on each slit while its imaginary part vanishes on the other parts of the real axis. Once we construct such a function F, then the desired function G will be defined as $G(z) = F(z) + \alpha z$ for some real constant α, as we shall see later. For the construction of F, we make use of the Weierstrass zeta function and the Schwarz–Christoffel formula.

For given positive real numbers c and d, let $t_1 = 2c$ and $t_2 = i2d$. Then the Weierstrass zeta function $\zeta(w)$ is defined by

$$\zeta(w) := \frac{1}{w} + \sum_{t \neq 0} \left(\frac{1}{w-t} + \frac{1}{t} + \frac{w}{t^2} \right), \tag{2.8}$$

where the sum is over all $t = n_1 t_1 + n_2 t_2$ with integers n_1 and n_2 not both zero. The function ζ has the periodicity properties

$$\zeta(w + t_1) = \zeta(w) + \eta_1, \quad \zeta(w + t_2) = \zeta(w) + i\eta_2, \tag{2.9}$$

where η_1 and η_2 are constants satisfying

$$d\eta_1 - c\eta_2 = \pi. \tag{2.10}$$

See [1]. For each $t = n_1 t_1 + n_2 t_2$, its conjugate $\bar{t} = n_1 t_1 - n_2 t_2$ is on the same lattice of points as t lies on. Thus one can easily see that

$$\overline{\zeta(\bar{w})} = \zeta(w), \tag{2.11}$$

and hence $\zeta(w)$ is real when w is real. We also have

$$\overline{\zeta(-\bar{w})} = -\zeta(w), \tag{2.12}$$

from which it follows that $\zeta(w)$ is purely imaginary when w is purely imaginary. It then follows from (2.9) that η_1 and η_2 are real.

Note that by (2.11)

$$\overline{\zeta(\bar{w} - 2id)} = \overline{\zeta(\overline{w + t_2})} = \zeta(w) + i\eta_2, \tag{2.13}$$

and hence

$$\zeta(w) - \overline{\zeta(\bar{w} - 2id)} = -i\eta_2. \tag{2.14}$$

Thus we deduce that if $w = u - id$ with u real, then

$$\Im\zeta(u - id) = -\frac{\eta_2}{2}. \tag{2.15}$$

Similarly, using the first identity in (2.9) and (2.12), one can see that if $w = -c + iv$ with v real, then

$$\Re\zeta(-c + iv) = -\frac{\eta_1}{2}, \tag{2.16}$$

and if $w = c + iv$ with v real, then

$$\Re\zeta(c + iv) = \frac{\eta_1}{2}. \tag{2.17}$$

We will also need the following lemma.

Lemma 2.1. *When $d > c$ the following inequality holds:*

$$\Im\zeta(\pm c + iv) \geq \frac{\eta_2}{2d}v, \quad -d < v < 0. \tag{2.18}$$

Proof. Note first that $\Im\zeta(-c + iv) = \Im\zeta(c + iv)$ because of the first identity in (2.9). By scaling we may assume that $2c = 1$. Put $2d = \tau$ to shorten notation, and note that $\tau > 1$. Let

$$h(v) := \Im\zeta(c + iv) - \frac{\eta_2}{2d}v, \quad -\frac{\tau}{2} < v < 0.$$

Since $h(0) = h(-\frac{\tau}{2}) = 0$, it suffices to show that h is concave in $(-\frac{\tau}{2}, 0)$. Observe that

$$h''(v) = -2\Im \sum_{n_1,n_2} \frac{1}{(\frac{1}{2} + i(v - n_2\tau) - n_1)^3}.$$

From the well-known identity (see [1])

$$\sum_{m=-\infty}^{\infty} \frac{1}{(z - m)^2} = \frac{\pi^2}{\sin^2 \pi z},$$

we have

$$\sum_{m=-\infty}^{\infty} \frac{1}{(z - m)^3} = \frac{\pi^3 \cos \pi z}{\sin^3 \pi z}.$$

Therefore, we get

$$\begin{aligned}
h''(v) &= -2\pi^3 \Im \sum_{n_2=-\infty}^{\infty} \frac{\cos\pi\left(\frac{1}{2}+i(v-n_2\tau)\right)}{\sin^3\pi\left(\frac{1}{2}+i(v-n_2\tau)\right)} \\
&= 2\pi^3 \sum_{n=-\infty}^{\infty} \frac{\sinh\pi(v-n\tau)}{\cosh^3\pi(v-n\tau)} \\
&= 2\pi^3 \frac{\sinh\pi v}{\cosh^3\pi v} \\
&\quad + 2\pi^3 \sum_{n=1}^{\infty} \frac{\sinh\pi(v-n\tau)\cosh^3\pi(v+n\tau)+\sinh\pi(v+n\tau)\cosh^3\pi(v-n\tau)}{\cosh^3\pi(v-n\tau)\cosh^3\pi(v+n\tau)}.
\end{aligned}$$

Straightforward but tedious computation yields

$$\begin{aligned}
&\sinh\pi(v-n\tau)\cosh^3\pi(v+n\tau)+\sinh\pi(v+n\tau)\cosh^3\pi(v-n\tau) \\
&= \frac{1}{2}\sinh 2\pi v\left[2+\cosh 2\pi v\cosh 2\pi n\tau-\cosh^2 2\pi n\tau\right],
\end{aligned}$$

and hence

$$h''(v) = 8\pi^3\sinh\pi v\left[\frac{1}{4\cosh^3\pi v}+\sum_{n=1}^{\infty}\frac{2+\cosh 2\pi v\cosh 2\pi n\tau-\cosh^2 2\pi n\tau}{\left(\cosh 2\pi v+\cosh 2\pi n\tau\right)^3}\right].$$

Since $v<0$, it is now enough to show that the quantity inside the bracket, which we call $I(v)$, is positive. Indeed, we have

$$I(v) > \frac{1}{4\cosh^3\pi v}-\sum_{n=1}^{\infty}\frac{1}{\cosh 2\pi v+\cosh 2\pi n\tau}.$$

Since $-\frac{\tau}{2}<v<0$ and $\tau>1$, we now have

$$\begin{aligned}
I(v) &> \frac{1}{4\cosh^3\frac{\pi\tau}{2}}-2\sum_{n=1}^{\infty}e^{-2\pi n\tau} \\
&= \frac{1}{4\cosh^3\frac{\pi\tau}{2}}-\frac{2}{1-e^{-2\pi\tau}}e^{-2\pi\tau} \\
&= 2\left[\left(e^{-\frac{1}{6}\pi\tau}+e^{-\frac{7}{6}\pi\tau}\right)^{-3}-\frac{1}{1-e^{-2\pi\tau}}\right]e^{-2\pi\tau} \\
&> 2\left[\left(e^{-\frac{1}{6}\pi}+e^{-\frac{7}{6}\pi}\right)^{-3}-\frac{1}{1-e^{-2\pi}}\right]e^{-2\pi\tau}>0.
\end{aligned}$$

This completes the proof. We remark that the inequality is proved not only when $d>c$, but also when τ is such that the second to last line in the above chain of inequalities is positive. □

For a positive real number β, define h by

$$h(w) := \beta\left(\zeta(w-id)-\frac{\eta_2}{2d}w+i\frac{\eta_2}{2}\right). \tag{2.19}$$

Then h is a meromorphic function with poles at $2n_1c+2in_2d+id$ and satisfies

$$\begin{cases}
\Re h(-c+iv) = -\beta c_0, \\
\Re h(c+iv) = \beta c_0, \\
\Im h(u) = 0, \\
\Im h(u+id) = 0
\end{cases} \tag{2.20}$$

for u and v real, where

$$c_0 = \frac{\eta_1}{2} - \frac{c\eta_2}{2d} = \frac{\pi}{2d} \tag{2.21}$$

because of (2.10). Since $\beta > 0$, we also have from (2.18)

$$\Im h(\pm c + iy) > 0, \quad 0 < y < d, \quad \text{when } d > c. \tag{2.22}$$

Since $\zeta(w) = \frac{1}{w} + O(1)$ as $w \to 0$, we have

$$h(w) = \frac{\beta}{w - id} + O(1) \quad \text{as } w \to id. \tag{2.23}$$

Restricting our attention to the rectangle $R = \{z = x + iy \mid -c < x < c,\ 0 < y < d\}$, we now construct a conformal mapping from the upper half of the complex plane onto R. To this end, it is natural to use the Schwarz–Christoffel formula.

For $b > a > 0$, let

$$\begin{aligned} g(z) :&= (z^2 - a^2)^{-1/2}(z^2 - b^2)^{-1/2} \\ &= (z + b)^{-1/2}(z + a)^{-1/2}(z - a)^{-1/2}(z - b)^{-1/2}, \end{aligned}$$

and define for z in the upper half plane

$$w = \Phi(z) := -\int_0^z g(\xi) d\xi. \tag{2.24}$$

The mapping Φ maps the upper half plane onto the rectangle $R = \{z = x + iy \mid -c < x < c,\ 0 < y < d\}$, where

$$c = \int_0^a \frac{dx}{\sqrt{(a^2 - x^2)(b^2 - x^2)}} \quad \text{and} \quad d = \int_a^b \frac{dx}{\sqrt{(x^2 - a^2)(b^2 - x^2)}}\ . \tag{2.25}$$

Note that the intervals $[-b, -a]$ and $[a, b]$ on the real axis get mapped onto the vertical sides $\{-c + iy \mid 0 \le y \le d\}$ and $\{c + iy \mid 0 \le y \le d\}$ of R, $[-a, a]$ onto the bottom of R, and $(-\infty, -b) \cup (b, \infty)$ into the top of R. The point ∞ is mapped to $w = id$, and

$$\Phi(z) = id + O\left(\frac{1}{|z|}\right) \quad \text{as } |z| \to \infty. \tag{2.26}$$

To see this, we have

$$\begin{aligned} \Phi(z) &= -\int_0^z g(\xi) d\xi = -\int_0^\infty g(\xi) d\xi + \int_z^\infty g(\xi) d\xi \\ &= id + \int_z^\infty g(\xi) d\xi = id + \int_z^\infty O(|\xi|^{-2}) d\xi \\ &= id + O(|z|^{-1}) \end{aligned}$$

as $|z| \to \infty$.

We now define F in the upper half of $\mathbb{C}$ by

$$F(z) := (h \circ \Phi)(z), \tag{2.27}$$

where Φ is defined by (2.24). It then follows from (2.20) that

$$\begin{cases} \Re F(x+i0) = -\beta c_0, & x \in (-b,-a), \\ \Re F(x+i0) = \beta c_0, & x \in (a,b), \\ \Im F(x+i0) = 0, & x \in (-\infty,-b) \cup (b,\infty) \cup (-a,a), \end{cases} \tag{2.28}$$

and from (2.22) that

$$\Im F(x+i0) \geq 0, \quad x \in (-b,-a) \cup (a,b). \tag{2.29}$$

It also follows from (2.23) and (2.26) that

$$F(z) = \beta z + O(1) \quad \text{as } |z| \to \infty. \tag{2.30}$$

Because of (2.28), F has an obvious extension as a holomorphic function in $\mathbb{C} \setminus ([-b,-a] \cup [a,b])$ satisfying

$$\overline{F(\overline{z})} = F(z). \tag{2.31}$$

For a positive real number α, define G by

$$G(z) := F(z) + \alpha z, \tag{2.32}$$

and then define curves C_j^+, $j = 1,2$, by

$$C_1^+ := \left\{ \lim_{y \to 0+} G(x+iy) \mid -b \leq x \leq -a \right\}, \tag{2.33}$$

$$C_2^+ := \left\{ \lim_{y \to 0+} G(x+iy) \mid a \leq x \leq b \right\}. \tag{2.34}$$

Observe from (2.29) that, at least when $d > c$ and $\beta > 0$, the curves C_j^+ (except the endpoints) lie in the upper half plane and their endpoints lie on the real axis. In fact, the endpoints of C_1^+ are

$$G(-b) = -\beta \frac{\pi}{2d} - \alpha b \quad \text{and} \quad G(-a) = -\beta \frac{\pi}{2d} - \alpha a, \tag{2.35}$$

and those of C_2^+ are

$$G(a) = \beta \frac{\pi}{2d} + \alpha a \quad \text{and} \quad G(b) = \beta \frac{\pi}{2d} + \alpha b. \tag{2.36}$$

The positivity of α is necessary to ensure that $G(b) > G(a)$. We now define C_j^- to be the reflection of C_j^+ about the real axis, i.e.,

$$C_j^- := \{ \overline{z} \mid z \in C_j^+ \}, \quad j = 1,2. \tag{2.37}$$

Assuming $d > c$, we then define the domain B_j to be the domain whose boundary is $C_j^\pm$ for $j = 1,2$. These domains are determined by the choice of the four parameters a, b, α, and β. However if we replace a,b by $k_1 a, k_1 b$, then the corresponding inclusions are just rescaled by the factor k_1. The reason is as follows.

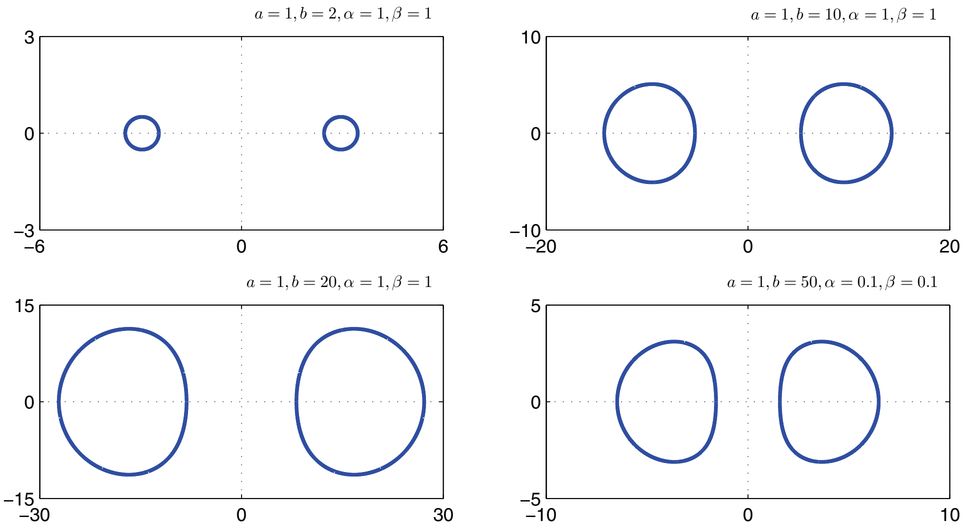


FIG. 2.1. *The typical shapes of inclusions. The scales for the figures are different. In the bottom right figure $\alpha = \beta = 0.1$, and the figure for $\alpha = \beta = 1$ is a similar one magnified ten times. In all figures, $c < d$.*

Let h_1, Φ_1, F_1, and G_1 be the functions defined by (2.19), (2.24), (2.27), (2.32), corresponding to $k_1 a, k_1 b$. Let h_0, etc., be those functions corresponding to a, b. Then we can see the following relations easily:

$$h_1(w) = k_1 h_0(k_1 w),$$
$$\Phi_1(z) = \frac{1}{k_1}\Phi_0\left(\frac{z}{k_1}\right).$$

Therefore, we have

$$G_1(z) = k_1 G_0\left(\frac{z}{k_1}\right).$$

This relation shows that the image of $[k_1 a, k_2 b]$ under G_1 is $k_1 C_2^+$, where C_2^+ is the image of $[a, b]$ under G_0 as given in (2.34).

If we replace α, β by $k_2\alpha, k_2\beta$, then the corresponding inclusions are just rescaled by k_2. This is more obvious. Thus without loss of generality one can choose $\alpha = \beta = 1$. If we just replace α by $k_3\alpha$, one can check that (2.28) implies that the boundary of each inclusion undergoes a linear stretching in the x-direction by a factor of k_3 (which is not in proportion to the change in the distance $2G(a)$ separating the inclusion pair). Thus, among all variations of the four parameters, changing only the ratio a/b leads to a nontrivial change in the inclusion shape. Figure 2.1 shows the shapes of B_1 and B_2, which are obtained numerically for various ratios a/b. Figure 2.2 shows a shape when $c > d$.

The following proposition shows that the inclusion constructed above enjoys the desired property.

PROPOSITION 2.2. *Let $B = B_1 \cup B_2$ be the inclusion constructed above. Then*

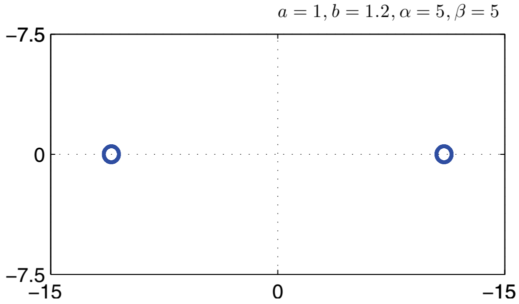


FIG. 2.2. *A shape when $c < d$. In this figure, $c = 1.7227$ and $d = 1.4310$.*

there is f holomorphic in $\mathbb{C} \setminus \overline{B}$ satisfying

$$f(z) = z + O(|z|^{-1}) \quad \textit{as } |z| \to \infty \tag{2.38}$$

and

$$f(z) = px + q_1 \quad \textit{for } z = x + iy \in \partial B_1, \tag{2.39}$$

$$f(z) = px + q_2 \quad \textit{for } z = x + iy \in \partial B_2, \tag{2.40}$$

for some real constant p and complex constants q_1 and q_2.

Proof. One can see from (2.33) and (2.34) that G is a homeomorphism from $[-b, -a] \cup [a, b]$ onto $C_1^+ \cup C_2^+$. One can also see that G is monotonically increasing on $(-\infty, -b]$, $[-a, a]$, and $[b, +\infty)$. Thus G is a homeomorphism from $\partial\Pi^+$ onto $\partial(\Pi^+ \setminus \overline{B_1 \cup B_2})$, where Π^+ is the complex upper half plane. Let φ and ψ be the conformal mappings from the unit disc Δ onto Π^+ and $\Pi^+ \setminus \overline{B_1 \cup B_2}$, respectively. Then $\psi^{-1} \circ G \circ \varphi : \Delta \to \Delta$ is holomorphic and a homeomorphism on $\partial\Delta$. Thus by Rado's theorem [27, p. 4], $\psi^{-1} \circ G \circ \varphi : \Delta \to \Delta$ is conformal, and hence univalent. Therefore, $G : \Pi^+ \to \Pi^+ \setminus \overline{B_1 \cup B_2}$ is univalent. Since $G(\bar{z}) = \overline{G(z)}$ and C_j^+ lies on the upper half plane, we conclude that G is univalent from $\mathbb{C} \setminus ([-b, -a] \cup [a, b])$ onto $\mathbb{C} \setminus \overline{B_1 \cup B_2}$. We emphasize that in order for G to be univalent, the upper part of ∂B_j, C_j^+ should lie on the upper half plane, as we proved before under the assumption that $d > c$. When $c < d$, the mapping G can sometimes be univalent and thus lead to other inclusion shapes, but we do not explore this possibility here.

Since G is univalent, G^{-1} is holomorphic in $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ and satisfies

$$G^{-1}(z) = \frac{1}{\alpha + \beta} z + O(1) \quad \text{as } |z| \to \infty, \tag{2.41}$$

and

$$G^{-1}(z) = \frac{1}{\alpha} x + \frac{\beta d}{2\pi\alpha} \quad \text{for } z = x + iy \in \partial B_1, \tag{2.42}$$

$$G^{-1}(z) = \frac{1}{\alpha} x - \frac{\beta d}{2\pi\alpha} \quad \text{for } z = x + iy \in \partial B_2. \tag{2.43}$$

Let

$$f(z) := (\alpha + \beta)[G^{-1}(z) - \gamma], \tag{2.44}$$

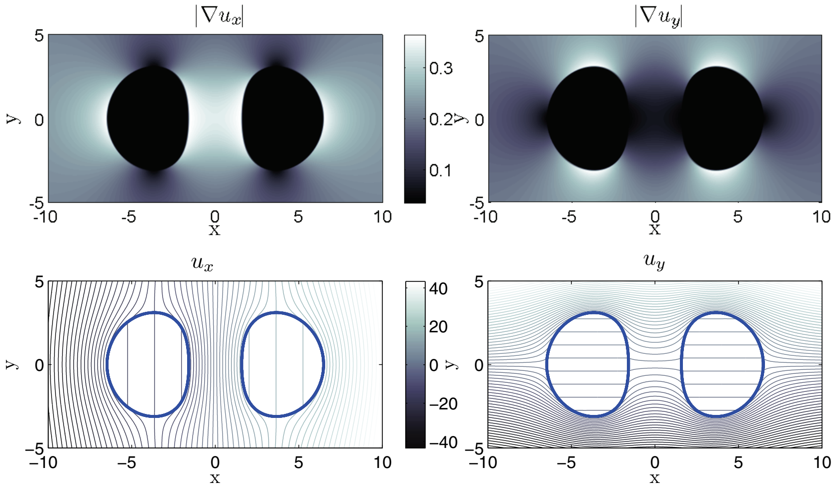


Fig. 3.1. *u_x and u_y are solutions corresponding to the $\mathbf{e}_1$ and $\mathbf{e}_2$ fields, respectively. The top figures are $|\nabla u_x|$ and $|\nabla u_y|$, and the bottom figures are equipotential lines of u_x and u_y.*

where γ is chosen so that f satisfies (2.38). By putting

$$p = \frac{\alpha+\beta}{\alpha}, \quad q_1 = (\alpha+\beta)\left[\frac{\beta d}{2\pi\alpha} - \gamma\right], \quad q_2 = (\alpha+\beta)\left[-\frac{\beta d}{2\pi\alpha} - \gamma\right], \tag{2.45}$$

we have (2.39) and (2.40). This completes the proof. □

We note that the most important property of f is that

$$\overline{-\frac{p}{2}z + f(z)} = \frac{p}{2}z + \overline{q_j} \quad \text{on } \partial B_j, \tag{2.46}$$

so that the function on the left-hand side of this equation, which is antiholomorphic outside the inclusion, can be extended inside B_j as a linear holomorphic function.

3. The uniformity property for antiplane elasticity.

We now show that the inclusions B_1 and B_2 have the uniformity property for antiplane elasticity (or for two-dimensional conductivity): For any uniform loading the field inside the inclusions is uniform. Before proving this, it may be helpful to the reader to refer to Figure 3.1, which clearly exhibits the uniformity property. This figure was obtained by solving (2.1) numerically using the boundary integral method.

We now prove the following theorem, which is a precise statement of the uniformity property for two-dimensional conductivity.

Theorem 3.1. *Let $B = B_1 \cup B_2$ be the inclusion constructed in section 2 with $\alpha > 0$ and $\beta > 0$. Let $k \neq 1$. For each nonzero constant vector a, let u be the solution to (2.1). Then ∇u is constant in B.*

Proof. Define U^e and U^i by

$$U^e(z) := \left(k + \frac{1-k}{p}\right)^{-1}\left(kz + \frac{1-k}{p}f(z)\right), \quad z \in \mathbb{C} \setminus \overline{B_1 \cup B_2}, \tag{3.1}$$

$$U^i(z) := \left(k + \frac{1-k}{p}\right)^{-1}(z + c_j), \quad z \in B_j, \ j = 1, 2, \tag{3.2}$$

where f is defined by (2.44). Choosing the complex constants c_j for $j = 1, 2$ properly, one can easily see that U^e and U^i satisfy (2.5) and $U^e(z) = z + O(|z|^{-1})$ as $|z| \to \infty$. Let $u := \Re U^e$ in $\mathbb{R}^2 \setminus \overline{B_1 \cup B_2}$ and $u := \Re U^i$ in $B_1 \cup B_2$. Then u satisfies (2.4), and hence u is the solution to (2.1) with $a = (1, 0)$. Note that we have

$$\nabla u = \frac{\alpha + \beta}{\alpha + k\beta} \mathbf{e}_1 \quad \text{in } B_j, \ j = 1, 2, \tag{3.3}$$

where $\mathbf{e}_1 = (1, 0)$. Thus for the uniform loading $\mathbf{e}_1 = (1, 0)$, the field inside B_1 and B_2 is given by (3.3).

One can show that the field inside the inclusion due to the loading $\mathbf{e}_2 = (0, 1)$ is also uniform using Keller's duality argument [19]. In fact, for a given $k \neq 1$, let $k_0 = 1/k$ and let u_0 be the solution to (2.1) with $a = \mathbf{e}_1$ and k replaced with k_0. Let v^e be the harmonic conjugate of u_0 in $\mathbb{C} \setminus \overline{B_1 \cup B_2}$ so that

$$v^e(x, y) - y = O(r^{-1}) \quad \text{as } r \to \infty. \tag{3.4}$$

The existence of such a harmonic conjugate is proved in [5]. Let v^i be the harmonic conjugate of u_0 in $B_1 \cup B_2$. Define w by

$$w(x, y) = \begin{cases} v^e(x, y), & (x, y) \in \mathbb{R}^2 \setminus \overline{B_1 \cup B_2}, \\ k_0 v^i(x, y) + C, & (x, y) \in B_1 \cup B_2, \end{cases} \tag{3.5}$$

where the constant C is chosen so that w is continuous across ∂B_j, $j = 1, 2$. Then using the Cauchy–Riemann equations one can show (see [5]) that w is the solution to (2.1) with $a = (0, 1)$. We also have from (3.3) and the Cauchy–Riemann equation that

$$\nabla w = \frac{\alpha + \beta}{k\alpha + \beta} \mathbf{e}_2 \quad \text{in } B_j, \ j = 1, 2. \tag{3.6}$$

This completes the proof. □

So far we have shown that the inclusions $B := B_1 \cup B_2$ have the uniformity property for the antiplane elasticity model: Given the applied field $\mathbf{e}_1$, the field inside B is uniform and given by $\frac{\alpha+\beta}{\alpha+k\beta}\mathbf{e}_1$, and for the applied field $\mathbf{e}_2$, the field inside B is uniform and given by $\frac{\alpha+\beta}{k\alpha+\beta}\mathbf{e}_2$.

4. Polarization tensors: Polyá–Szegö conjecture. In this section we compute the polarization tensor associated with $B = B_1 \cup B_2$ and show that the Polyá–Szegö conjecture fails to be true among inclusions with multiple components. To explain the polarization tensor associated with the inclusion B consisting of m components $B_1, \ldots, B_m$, we consider the following problem: For a vector $\xi \in \mathbb{R}^2$,

$$\begin{cases} \nabla \cdot \big(1 + (k-1)\chi(B)\big)\nabla u = 0 \quad \text{in } \mathbb{R}^2, \\ u(x, y) - \xi \cdot (x, y) = O(r^{-1}) \quad \text{as } r \to \infty. \end{cases} \tag{4.1}$$

The solution u to (4.1) admits the asymptotic expansion

$$u(x, y) = \xi \cdot (x, y) + \frac{1}{2\pi} \xi \cdot M \frac{(x, y)^T}{r^2} + O(r^{-2}) \quad \text{as } r \to \infty \tag{4.2}$$

for some 2×2 matrix M. This matrix $M = M(B)$ is the polarization tensor associated with B. It should be noted that the polarization tensor associated with the inclusion consisting of multiple components $B_1, \dots, B_m$ is not the sum or a combination of the polarization tensors of individual inclusions. It incorporates the interactions among components.

It is known that the eigenvalues of the polarization tensor must be confined within the so-called Hashin–Shtrikman bounds [21, 7] (see also [20, 23]):

$$\mathrm{Tr}(M) \leq (k-1)\left(1+\frac{1}{k}\right)|B| \tag{4.3}$$

and

$$\mathrm{Tr}(M^{-1}) \leq \frac{1+k}{(k-1)|B|}, \tag{4.4}$$

where Tr denotes the trace and $|B|$ is the area of B. If M has minimal trace, then M satisfies (4.4) and M is diagonal. These bounds are known to be optimal in the sense that all the points inside the bound, except the upper bound, are realized as the pair of eigenvalues of the polarization tensor associated with a certain shape—coated ellipses [7] and crosses [3]. The lower bound (4.4) is attained by ellipses. Thus a conjecture, which implies the Polyá–Szegö conjecture, is that if (4.4) holds for an inclusion, then that inclusion must be an ellipse.

Kang and Milton [18] proved this new conjecture affirmatively in two dimensions (and three dimensions) within the class of simply connected inclusions with Lipschitz boundaries. In fact, in [17], they showed that if the polarization tensor $M(B)$ satisfies the lower bound (4.4), then B must have the uniformity property and is therefore an ellipse by Eshelby's conjecture. The Pólya–Szegö conjecture, which asserts that the inclusion whose polarization tensor has the minimal trace is a disk, follows from this.

We now show that the polarization tensor associated with the inclusion constructed in section 2 satisfies (4.4), and hence the Pólya–Szegö conjecture is not true among nonsimply connected inclusions. To do that, let u^1 and u^2 be solutions to (2.1) with $a = \mathbf{e}_1$ and $a = \mathbf{e}_2$, respectively, and put $\mathbf{u} := (u^1, u^2)$. Then, the polarization tensor M is given by

$$M = (k-1)\int_B \nabla \mathbf{u}\, dxdy, \tag{4.5}$$

where $\nabla \mathbf{u}$ is the Jacobian matrix. See [4], for example, for the proof of (4.5). As an immediate consequence of (3.3) and (3.6) we obtain the following corollary.

Corollary 4.1. *The polarization tensor associated with the inclusion constructed in section* 2 *is given by*

$$M = (k-1)|B| \begin{pmatrix} \dfrac{\alpha+\beta}{\alpha+k\beta} & 0 \\ 0 & \dfrac{\alpha+\beta}{k\alpha+\beta} \end{pmatrix}. \tag{4.6}$$

Note that this tensor satisfies

$$\mathrm{Tr}(M^{-1}) = \frac{k+1}{(k-1)|B|}, \tag{4.7}$$

which is the lower Hashin–Shtrikman bound. It is quite interesting to observe that the polarization tensor (4.6) is the same as that for the ellipse $\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} \leq 1$. In particular, the inclusion in Figure 2.1, which has $\alpha = \beta$, has the same polarization tensor as that of a circular disk.

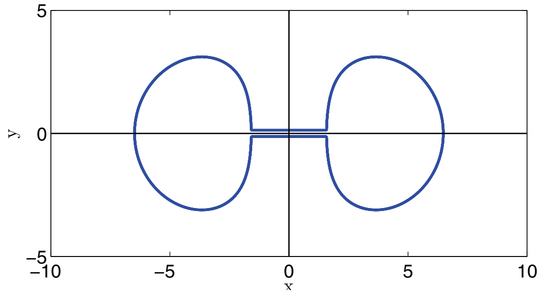


FIG. 5.1. *Bridged inclusion.*

5. Instability of the uniformity property. For a given $\epsilon > 0$, let δ be a positive number such that the rectangle $(-\delta, \delta) \times (\epsilon, \epsilon)$ is contained in the convex hull of B_1 and B_2. Let $B_\epsilon := B \cup ((-\delta, \delta) \times (\epsilon, \epsilon))$. B_ϵ is B_1 and B_2 connected by a thin bridge. Figure 5.1 shows the bridged inclusion.

Let γ and γ_ϵ be the conductivity distributions with inclusions B and B_ϵ, respectively, namely,

$$\gamma = 1 + (k-1)\chi(B), \quad \gamma_\epsilon = 1 + (k-1)\chi(B_\epsilon). \tag{5.1}$$

Let $h(x, y)$ be a harmonic function in $\mathbb{R}^2$, e.g., $h(x, y) = x$ or y. Let u be the solution to

$$\begin{cases} \nabla \cdot \gamma \nabla u = 0 \quad \text{in } \mathbb{R}^2, \\ u(x,y) - h(x,y) = O(r^{-1}) \quad \text{as } r \to \infty \end{cases} \tag{5.2}$$

and u_ϵ be the solution to (5.2) with γ replaced with γ_ϵ. Then a standard regularity theory of elliptic equations shows that

$$\|\nabla(u - u_\epsilon)\|_2 \to 0 \quad \text{as } \epsilon \to 0. \tag{5.3}$$

Here $\| \cdot \|_2$ is the norm of the square integral. In fact, if we put $w = u - u_\epsilon$, then w satisfies

$$\begin{cases} \nabla \cdot \gamma_\epsilon \nabla w = \nabla \cdot (\gamma_\epsilon - \gamma) \nabla u \quad \text{in } \mathbb{R}^2, \\ w(x,y) = O(r^{-1}) \quad \text{as } r \to \infty. \end{cases} \tag{5.4}$$

Thus it follows from a regularity theorem for the elliptic operator $\nabla \cdot \gamma_\epsilon \nabla$ (see [11]) that provided k is strictly positive

$$\|w\|_{H^1(\mathbb{R}^2)} \leq C \left(\int_{R_\epsilon} |\nabla u|^2 \right)^{1/2} \tag{5.5}$$

for some constant C independent of ϵ, where $R_\epsilon = B_\epsilon \setminus B$. In particular, we have (5.3).

If $h(x, y) = x$ or y, ∇u is constant in B_1 and B_2, as we have seen in section 3. Therefore, by (5.3), ∇u_ϵ is almost uniform (in the H^1 sense) if ϵ is small. It is obvious that B_ϵ is simply connected but nothing similar to an ellipse. Figure 5.2 shows the absolute value of the gradient of u_ϵ.

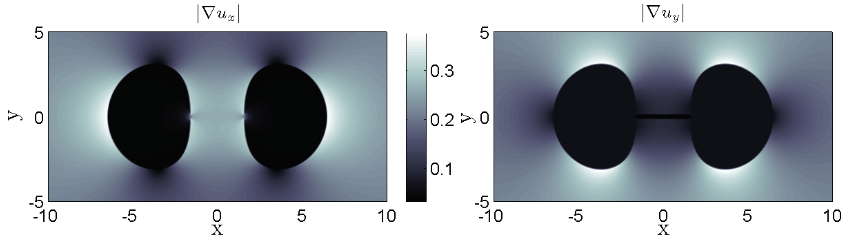


FIG. 5.2. *The graph of the absolute value of the gradient of the solutions corresponding to the bridged inclusion.*

We note that (5.5) implies that

$$\|M(B) - M(B_\epsilon)\| \leq C\epsilon, \tag{5.6}$$

where $M(B)$ is the polarization tensor of B and $M(B_\epsilon)$ is that of B_ϵ. In the case when $\alpha = \beta$, this equation shows that from a practical standpoint the Polyá–Szegö conjecture is false in two dimensions: a simply connected inclusion can have a polarizability tensor arbitrarily close to that of a circular disk yet not resemble a disk at all. We remark that in the extreme cases not treated here, when $k = 0$ or $k = \infty$, the insertion of even an infinitesimal bridge drastically changes the polarization tensor. So it is still an open question whether a void or perfectly conducting region is necessarily close in shape to an ellipse if it is simply connected and almost has the polarizability tensor of an ellipse.

6. Uniformity property: The elasticity case.

In this section we consider the uniformity property of the inclusion $B_1 \cup B_2$ for planar elasticity and show that for a certain loading the field inside $B_1 \cup B_2$ is uniform while for other loadings it is not uniform.

Let $C = (C_{ijkl})$ be the elasticity tensor of the inclusion-matrix composite, namely,

$$C_{ijkl} := \Big(\lambda\, \chi(\mathbb{R}^2 \setminus \overline{B}) + \widetilde{\lambda}\, \chi(B)\Big)\delta_{ij}\delta_{kl} + \Big(\mu\, \chi(\mathbb{R}^2 \setminus \overline{B}) + \widetilde{\mu}\, \chi(B)\Big)(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}),$$

where $B = B_1 \cup B_2$. The elasticity tensor C indicates that the matrix (the background) has Lamé parameters (λ, μ), while the inclusion has parameters $(\widetilde{\lambda}, \widetilde{\mu})$. It is always assumed that

$$\mu > 0, \quad d\lambda + 2\mu > 0, \quad \widetilde{\mu} > 0, \quad \text{and} \quad d\widetilde{\lambda} + 2\widetilde{\mu} > 0$$

for ellipticity. For given constants a_{ij}, $i, j = 1, 2$, consider the following linear elastic problem:

$$\begin{cases} \nabla \cdot \big(C(\nabla \mathbf{u} + \nabla \mathbf{u}^T)\big) = 0 & \text{in } \mathbb{R}^2, \\ \mathbf{u}(x) - \displaystyle\sum_{i,j=1}^{2} a_{ij} x_i \mathbf{e}_j = O(|x|^{-1}) & \text{as } |x| \to \infty, \end{cases} \tag{6.1}$$

where $\mathbf{e}_j$, $j = 1, \ldots, d$, denotes the standard basis for $\mathbb{R}^2$. The uniform applied loading is determined by the matrix (a_{ij}).

Let us first seek a type of loading which yields a uniform field inside the inclusions. The existence of such a loading is expected due to the link between conductivity problems and elasticity problems in composites when the field in one phase is uniform [13]. We first invoke the following complex representation of the solution to (6.1) from [24, 4]: Let $\mathbf{u} = (u, v)$ be the solution of (6.1) and let $\mathbf{u}_e := \mathbf{u}|_{\mathbb{C}\setminus\overline{B}}$ and $\mathbf{u}_i := \mathbf{u}|_B$. Then there are unique functions φ_e and ψ_e holomorphic in $\mathbb{C} \setminus \overline{B}$ and φ_i and ψ_i holomorphic in B such that

$$2\mu(u_e + iv_e)(z) = \kappa\varphi_e(z) - z\overline{\varphi_e'(z)} - \overline{\psi_e(z)}, \quad z \in \mathbb{C} \setminus \overline{B}, \tag{6.2}$$

$$2\widetilde{\mu}(u_i + iv_i)(z) = \widetilde{\kappa}\varphi_i(z) - z\overline{\varphi_i'(z)} - \overline{\psi_i(z)}, \quad z \in B, \tag{6.3}$$

where

$$\kappa = \frac{\lambda + 3\mu}{\lambda + \mu}, \qquad \widetilde{\kappa} = \frac{\widetilde{\lambda} + 3\widetilde{\mu}}{\widetilde{\lambda} + \widetilde{\mu}}. \tag{6.4}$$

Moreover, the following hold on ∂B_j, $j = 1, 2$:

$$\frac{1}{2\mu}\left(\kappa\varphi_e(z) - z\overline{\varphi_e'(z)} - \overline{\psi_e(z)}\right) = \frac{1}{2\widetilde{\mu}}\left(\widetilde{\kappa}\varphi_i(z) - z\overline{\varphi_i'(z)} - \overline{\psi_i(z)}\right), \tag{6.5}$$

$$\varphi_e(z) + z\overline{\varphi_e'(z)} + \overline{\psi_e(z)} = \varphi_i(z) + z\overline{\varphi_i'(z)} + \overline{\psi_i(z)} + c, \tag{6.6}$$

where c is a constant. Equation (6.5) expresses continuity of displacement, and (6.6) expresses continuity of traction.

Let f be the function in (2.44) and let

$$\varphi_e(z) = A_e z, \quad \psi_e(z) = C_e\left[-\frac{p}{2}z + f(z)\right], \quad z \in \mathbb{C} \setminus \overline{B}, \tag{6.7}$$

where A_e and C_e are complex and real constants, respectively. As was observed in (2.46), $\overline{\psi}$ on ∂B_j has an extension to B_j as the linear holomorphic function $C_e(\frac{p}{2}z + \overline{q_j})$. Therefore, on ∂B_j, $j = 1, 2$, (6.5) and (6.6) now take the forms

$$\left(\widetilde{\kappa}\varphi_i(z) - z\overline{\varphi_i'(z)} - \overline{\psi_i(z)}\right) = \frac{\widetilde{\mu}}{\mu}\left(\kappa A_e - \overline{A_e} - \frac{C_e p}{2}\right)z + D_j, \tag{6.8}$$

$$\varphi_i(z) + z\overline{\varphi_i'(z)} + \overline{\psi_i(z)} = \left(A_e + \overline{A_e} + \frac{C_e p}{2}\right)z + E_j \tag{6.9}$$

for some constants D_j and E_j. Equations (6.8) and (6.9) force us to take $\varphi_i(z) = A_i z + \text{constant}$ and $\psi_i = \text{constant}$, and the complex number A_i should satisfy

$$\begin{cases} \widetilde{\kappa}A_i - \overline{A_i} = \dfrac{\widetilde{\mu}}{\mu}\left(\kappa A_e - \overline{A_e} - \dfrac{C_e p}{2}\right), \\[2ex] A_i + \overline{A_i} = A_e + \overline{A_e} + \dfrac{C_e p}{2}. \end{cases} \tag{6.10}$$

Let $A_e = a_1 + ia_2$. Equation (6.10) has a solution A_i if and only if

$$C_e = \frac{4}{p}\left[1 + \frac{2\widetilde{\mu}}{\mu(\widetilde{\kappa} - 1)}\right]^{-1}\left[\frac{\widetilde{\mu}(\kappa - 1)}{\mu(\widetilde{\kappa} - 1)} - 1\right]a_1, \tag{6.11}$$

and in this case

$$2\mu\begin{pmatrix} u_e \\ v_e \end{pmatrix} = \begin{pmatrix} (\kappa - 1)a_1 - C_e(1 - \frac{p}{2}) & -(\kappa + 1)a_2 \\ (\kappa + 1)a_2 & (\kappa - 1)a_1 + C_e(1 - \frac{p}{2}) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + O(r^{-1}) \tag{6.12}$$

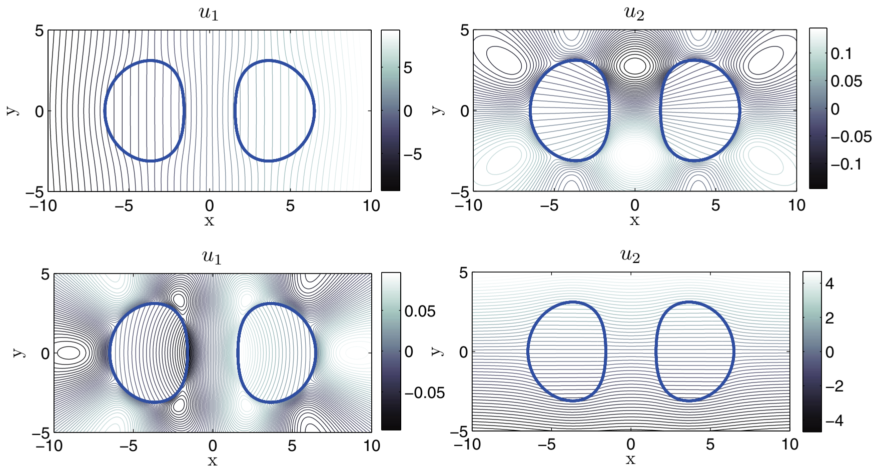


FIG. 6.1. *Equipotential lines of the solution* $\mathbf{u} = (u_1, u_2)$ *for the loading* $(x, 0)$ *(top) and* $(0, y)$ *(bottom).* ∇u_2 *is not uniform for the top, while* ∇u_1 *is not uniform for the bottom.*

as $r \to \infty$. Putting $t = \frac{(\kappa-1)a_1}{2\mu}$ and $s = \frac{(\kappa+1)a_2}{2\mu}$ and simplifying expressions using (2.45) and (6.4), we arrive at the following: If the loading (a_{ij}) is of the form

$$(a_{ij}) = \begin{pmatrix} (1-\theta)t & -s \\ s & (1+\theta)t \end{pmatrix}, \tag{6.13}$$

where t and s are real constants and

$$\theta = \frac{(\alpha - \beta)(\widetilde{\lambda} + \widetilde{\mu} - \lambda - \mu)}{(\alpha + \beta)(\mu + \widetilde{\lambda} + \widetilde{\mu})}, \tag{6.14}$$

then $\nabla \mathbf{u}$ is constant in B where $\mathbf{u}$ is the solution to (6.1). In particular, when $\alpha = \beta$, this corresponds to a hydrostatic loading. We mention that the inclusions constructed in this paper depend on the parameters α and β. In summary, we have proved the following theorem.

THEOREM 6.1. *If the* (a_{ij}) *are given by* (6.13) *for some real numbers* s *and* t *where* θ *is defined by* (6.14)*, then the solution* $\mathbf{u}$ *to* (6.1) *has the property that* $\nabla \mathbf{u}$ *is constant in* B*.*

We do not have a complete characterization of those loadings which yield a uniform strain field inside the inclusion, but numerical computations show that for certain loadings the field is not uniform. Figure 6.1 shows the equipotential lines for the solution $\mathbf{u} = (u_1, u_2)$ for the loadings $(x, 0)$ and $(0, y)$. It is worthwhile to compare the result of this paper with that for the simply connected inclusion in [28, 18]. For a simply connected inclusion, if the field inside the inclusion is uniform for a single loading, then the inclusion is of elliptical shape, and hence the field is uniform for any loading. Here we established that it is not the case for an inclusion with multiple components. It is an open question whether the uniformity of the interior field for all uniform applied loadings forces the inclusion (with possibly multiple components) to be an ellipse or not.

Conclusion. Providing an alternative proof to that of Cherepanov [8], we constructed a family of inclusions with two components which have the uniformity property for antiplane elasticity: for any loading the field inside the inclusions is uniform. In the case of planar elasticity the field is uniform for certain types of loadings but not uniform for other loadings. These results show that the conjectures of Eshelby and Pólya–Szegö are not true among nonsimply connected inclusions. By connecting two inclusions by a thin bridge we showed that these conjectures do not hold in a practical sense even for simply connected inclusions: even if the field inside an inclusion is very close to being uniform, the inclusion need not be close to an ellipse.

Acknowledgments. The authors are grateful to the referees for comments and in particular to one referee for drawing our attention to the work of Cherepanov. Also, after this work was submitted, the authors became aware of the beautiful paper of Liu [22], which treats multiple inclusions with the Eshelby property not only in two dimensions but also in three dimensions.

REFERENCES

[1] L. V. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1978.

[2] G. Alessandrini and V. Isakov, *Analyticity and uniqueness for the inverse conductivity problem*, Rend. Istit. Mat. Univ. Trieste, 28 (1996), pp. 351–369.

[3] H. Ammari, Y. Capdeboscq, H. Kang, E. Kim, and M. Lim, *Attainability by simply connected domains of optimal bounds for polarization tensors*, European J. Appl. Math., 17 (2006), pp. 201–219.

[4] H. Ammari and H. Kang, *Reconstruction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, New York, 2004.

[5] H. Ammari, H. Kang, and M. Lim, *Gradient estimates for solutions to the conductivity problem*, Math. Ann., 332 (2005), pp. 277–286.

[6] K. Astala and V. Nesi, *Composites and quasiconformal mappings: New optimal bounds in two dimensions*, Calc. Var. Partial Differential Equations, 18 (2003), pp. 335–355.

[7] Y. Capdeboscq and M. S. Vogelius, *A review of some recent work on impedance imaging for inhomogeneities of low volume fraction*, in Partial Differential Equations and Inverse Problems, Contemp. Math. 362, AMS, Providence, RI, 2004, pp. 69–87.

[8] G. P. Cherepanov, *Inverse problems of the plane theory of elasticity*, Prikl. Mat. Meh., 38 (1974), pp. 963–979 (in Russian).

[9] J. D. Eshelby, *The determination of the elastic field of an ellipsoidal inclusion, and related problems*, Proc. Roy. Soc. London Ser. A, 241 (1957), pp. 376–396.

[10] J. D. Eshelby, *Elastic inclusions and inhomogeneities*, in Progress in Solid Mechanics, Vol. II, North–Holland, Amsterdam, 1961, pp. 87–140.

[11] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.

[12] Y. Grabovsky and R. Kohn, *Microstructures minimizing the energy of a two phase elastic composite in two space dimensions.* II. *The Vigdergauz microstructure*, J. Mech. Phys. Solids, 43 (1995), pp. 949–972.

[13] Y. Grabovsky, *Bounds and extremal microstructures for two-component composites: A unified treatment based on the translation method*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 919–944.

[14] Z. Hashin, *The elastic moduli of heterogeneous materials*, Trans. ASME Ser. E J. Appl. Mech., 29 (1962), pp. 143–150.

[15] C. O. Horgan, *Anti-plane shear deformations in linear and nonlinear solid mechanics*, SIAM Rev., 37 (1995), pp. 53–81.

[16] V. Isakov and J. Powell, *On the inverse conductivity problem with one measurement*, Inverse Problems, 6 (1990), pp. 311–318.

[17] H. Kang and G. W. Milton, *On Conjectures of Pólya–Szegö and Eshelby*, in Inverse Problems, Multi-scale Analysis and Effective Medium Theory, Contemp. Math. 408, AMS, Providence, RI, 2006, pp. 75–80.

[18] H. Kang and G. W. Milton, *Solutions to the Pólya–Szegö conjecture and the weak Eshelby conjecture*, Arch. Ration. Mech. Anal., 188 (2008), pp. 93–116.

[19] J. B. Keller, *A theorem on the conductivity of a composite medium*, J. Math. Phys., 5 (1964), pp. 548–549.

[20] R. V. Kohn and G. W. Milton, *On bounding the effective conductivity of anisotropic composites*, in Homogenization and Effective Moduli of Materials and Media, J. L. Ericksen, D. Kinderlehrer, R. V. Kohn, and J. L. Lions, eds., IMA Vol. Math. Appl. 1, Springer-Verlag, New York, 1986, pp. 97–125.

[21] R. Lipton, *Inequalities for electric and elastic polarization tensors with applications to random composites*, J. Mech. Phys. Solids, 41 (1993), pp. 809–833.

[22] L. P. Liu, *Solutions to the Eshelby conjectures*, Proc. Roy. Soc. A Math. Phys. Engrg. Sci., 464 (2008), pp. 573–594.

[23] G. W. Milton, *The Theory of Composites*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2002.

[24] N. I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*, English translation, Noordhoff International Publishing, Leiden, 1977.

[25] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.

[26] C.-Q. Ru and P. Schiavone, *On the elliptic inclusion in anti-plane shear*, Math. Mech. Solids, 1 (1996), pp. 327–333.

[27] R. Schoen and S. T. Yau, *Lectures on Harmonic Maps*, International Press, Boston, 1997.

[28] G. P. Sendeckyj, *Elastic inclusion problems in plane elastostatics*, Internat. J. Solids Structures, 6 (1970), pp. 1535–1543.

[29] S. B. Vigdergauz, *Effective elastic parameters of a plate with a regular system of equal-strength holes*, Inzhenernyi Zhurnal. Mekhnika Tverdogo Tela, 21 (1986), pp. 165–169.

[30] S. B. Vigdergauz, *Two dimensional grained composites of extreme rigidity*, J. Appl. Mech., 61 (1994), pp. 390–394.

[31] L. T. Wheeler, *Stress minimum forms for elastic solids*, AMR, 45 (1992), pp. 1–12.

SIAM J. APPL. MATH.
Vol. 69, No. 2, pp. 596–620

QUALITATIVE ANALYSIS OF A PREY-PREDATOR MODEL WITH STAGE STRUCTURE FOR THE PREDATOR[*]

YIHONG DU[†], PETER Y. H. PANG[‡], AND MINGXIN WANG[§]

Abstract. In this paper, we propose a diffusive prey-predator model with stage structure for the predator. We first analyze the stability of the nonnegative steady states for the reduced ODE system and then study the same question for the corresponding reaction-diffusion system with homogeneous Neumann boundary conditions. We find that a Hopf bifurcation occurs in the ODE system, but no Turing pattern happens in the reaction-diffusion system. However, when a natural cross diffusion term is included in the model, we can prove the emergence of stationary patterns (i.e., nonconstant positive stationary solutions) for this system; moreover, these stationary patterns do not exist in the considered parameter regime when there is no cross diffusion.

Key words. predator-prey model, stage structure, stability, cross diffusion, Turing pattern

AMS subject classifications. 35J55, 92D25

DOI. 10.1137/070684173

1. Introduction and the mathematical model.

The life histories of plants, insects, and animals exhibit enormous diversity. Most species go through several stages during their lifetime, such as immature and mature stages; more specialized stages may exist for dispersal or dormancy. The vital rates (rates of survival, development, and reproduction) almost always depend on the development stage, among many other factors. Such stage structures have been largely ignored in early population modeling but have received much attention in recent years; see [2, 9, 11, 16, 22, 23, 31, 41, 42, 44, 45] and the references therein. Generally speaking, population growth models that include stage structure predict more complex population dynamics than those without taking this factor into account.

Hitherto, a common assumption in stage-structured prey-predator models is that the immature predator has no direct effect on the prey or the mature predator, and the immature prey is not subject to predation.

For example, in [22], the authors studied the following predator-prey models with

[*] Received by the editors March 2, 2007; accepted for publication (in revised form) August 5, 2008; published electronically December 3, 2008.

http://www.siam.org/journals/siap/69-2/68417.html

[†] School of Mathematics, Statistics and Computer Science, University of New England, Armidale, NSW 2351, Australia; and Department of Mathematics, Qufu Normal University, Qufu, Shandong 273165, People's Republic of China (ydu@turing.une.edu.au). The work of this author was supported by the Australian Research Council.

[‡] Department of Mathematics, National University of Singapore, 2 Science Drive 2, Republic of Singapore 117543 (matpyh@nus.edu.sg). The work of this author was supported by the NUS ARF grants R-146-000-055-112 and R-146-000-088-112.

[§] Department of Mathematics, Southeast University, Nanjing 210018, People's Republic of China; and Department of Mathematics, Xuzhou Normal University, Xuzhou 221116, People's Republic of China (mxwang@seu.edu.cn). The work of this author was supported by the National Natural Science Foundation of China grants 10471022 and 10771032, the Natural Science Foundation of Jiangsu Province grant BK2006088, and the Ministry of Education of China Science and Technology Major Projects grant 104090. This paper was completed during visits of the third author to the National University of Singapore and the University of New England which were partially supported by the NUS ARF grants R-146-000-055-112 and R-146-000-088-112 and the Australian Research Council.

stage structure for the predator:

$$
(1.1)\quad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = rx(t)\left(1 - \dfrac{x(t)}{K}\right) - \dfrac{bx(t)y(t)}{1 + k_1x(t) + k_2y(t)}, \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = \dfrac{nb\mathrm{e}^{-d_j\tau}x(t-\tau)y(t-\tau)}{1 + k_1x(t-\tau) + k_2y(t-\tau)} - dy(t), \\ \dfrac{\mathrm{d}y_j}{\mathrm{d}t} = \dfrac{nbx(t)y(t)}{1 + k_1x(t) + k_2y(t)} - \dfrac{nb\mathrm{e}^{-d_j\tau}x(t-\tau)y(t-\tau)}{1 + k_1x(t-\tau) + k_2y(t-\tau)} - d_jy_j(t), \end{cases}
$$

where x, y, and y_j are the densities of the prey, mature predator, and immature predator, respectively, d_j is the through-stage death rate of the immature predator, and the time delay τ is the duration that each immature predator needs to reach maturity. Thus $\mathrm{e}^{-d_j\tau}$ is the surviving rate of the immature predator (to mature predator). We note that the first two equations in this model are independent of the immature predator y_j, and they completely determine the dynamics of the system.

In this paper we propose a diffusive prey-predator model with stage structure for the predator which includes an explicit interaction between the immature predator and the prey as well as the mature predator. In particular, the interaction between the immature predator and the mature predator gives rise to a cross diffusion term (see (1.6) below). The resulting mathematical model is a strongly coupled system of three equations which is mathematically much more complex than those considered earlier. In particular, we will demonstrate several special features of the model which cannot be captured by the corresponding ODE model or by the corresponding reaction-diffusion model without cross diffusion.

To describe our model, we start with the classical Lotka–Volterra ODE model. Let x and y be the densities of the prey and the predator, respectively. We divide the predator y into two parts, the immature y_1 and mature y_2, with $y = y_1 + y_2$. Taking into account several biological considerations not adequately addressed before and with the view of simplicity, we make the following assumptions:

(i) The immature predators do not yield offspring.

(ii) The interaction terms are of Lotka–Volterra type, i.e., based on linear functional response.

(iii) The predation rate of the immature predator (which could be indirect, through increased consumption of prey by the mothers, for example) is positive but less than that of the mature predator. We denote them by εC and C, respectively, with $0 < \varepsilon < 1$.

(iv) The rate of conversion of prey to immature predator is proportional to the prey and is denoted by Kx. (In some literature, the constant K/C is called the rate of conversion of nutrients into the production of the immature predator.) The death rate of the immature predator is a constant denoted by M.

(v) In general, the rate of transition from immature predator to mature predator is a function of the prey. For simplicity, we assume that this rate is a constant denoted by D. The death rate of the mature predator is a constant denoted by P.

Under the above assumptions, the ODE prey-predator model with stage structure for the predator can be written as

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = Ax - Bx^2 - \varepsilon Cxy_1 - Cxy_2, & t > 0, \\ \dfrac{\mathrm{d}y_1}{\mathrm{d}t} = Kxy_2 - Dy_1 - My_1, & t > 0, \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}t} = Dy_1 - Py_2, & t > 0. \end{cases} \tag{1.2}$$

Using the scaling $u = Bx/(M + D)$, $v = Cy_1/(M + D)$, $w = Cy_2/(M + D)$, $\tau = (M + D)t$, and denoting τ by t again, the system (1.2) becomes

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = au - u^2 - \varepsilon uv - uw, & t > 0, \\ \dfrac{\mathrm{d}v}{\mathrm{d}t} = kuw - v, & t > 0, \\ \dfrac{\mathrm{d}w}{\mathrm{d}t} = bv - mw, & t > 0, \end{cases} \tag{1.3}$$

where $a = A/(M + D)$, $b = D/(M + D) < 1$, $k = K/B$, and $m = P/(M + D)$.

Two obvious nonnegative steady states of the system (1.3) are $\mathbf{u}_0 = (0, 0, 0)$ and $\mathbf{u}_a = (a, 0, 0)$. Moreover, the problem (1.3) has a positive steady state if and only if $m < abk$, in which case the positive steady state is uniquely given by $\tilde{\mathbf{u}} = (\tilde{u}, \tilde{v}, \tilde{w})$, where

$$\tilde{u} = \frac{m}{bk}, \quad \tilde{v} = \frac{m(abk - m)}{bk(b + m\varepsilon)}, \quad \tilde{w} = \frac{abk - m}{k(b + m\varepsilon)}. \tag{1.4}$$

Now, the ODE model reflects only population changes due to predation in a situation where predator and prey densities are not spatially dependent. It does not take into account the fact that population is usually not homogeneously distributed, or the fact that the movements of the predator and prey are also caused by interactions within the same species or with other species. These considerations may be modeled by diffusion processes which can be quite intricate as different concentration levels of prey and predator cause different population movements. Such movements can be determined by the concentration of the same species (diffusion) and that of the other species (cross diffusion).

The role of diffusion in the modeling of many physical, chemical, and biological processes has been extensively studied. Starting with Turing's seminal 1952 paper [37], diffusion has been observed as causes of the spontaneous emergence of ordered structures, called patterns, in a variety of nonequilibrium situations. More recently, cross diffusion, in addition to diffusion, has also been used in some of these problems. There is a great variety of models that involve the applications of diffusion and cross diffusion; some examples are the Gierer–Meinhardt model [14, 18, 46], the Sel'kov model [20, 38], the chemotaxis diffusion model [21, 43], the competition model [24, 25, 26], the predator-prey model [7, 13, 19, 33, 34, 35, 40], and models of semiconductors, plasmas, chemical waves, combustion systems, embryogenesis, etc.; see, e.g., [3, 5, 10, 36] and references therein.

A stable stationary solution for an ODE system may lose its stability when regarded as a stationary solution of the corresponding reaction-diffusion system (i.e., with diffusion added to the system) over a bounded domain with Neumann boundary conditions and induce space dependent stationary solutions through local bifurcation. Turing [37] first observed this phenomenon and suggested that this may be used to model pattern formation. Such diffusion-driven instability, now known as Turing instability, has been verified empirically [4, 32]. Patterns arising from such a situation are called Turing patterns.

For our model, taking into account the inhomogeneous distribution of the prey and the predator in different spatial locations within a fixed bounded domain Ω at any given time, and the natural tendency of each species to diffuse to areas of smaller population concentration, we are naturally led to the following corresponding reaction-diffusion system:

$$(1.5)\qquad \begin{cases} u_t - d_1\Delta u = au - u^2 - \varepsilon uv - uw, & x\in\Omega,\ t>0,\\ v_t - d_2\Delta v = kuw - v, & x\in\Omega,\ t>0,\\ w_t - d_3\Delta w = bv - mw, & x\in\Omega,\ t>0,\\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x\in\partial\Omega,\ t>0,\\ w(x,0)\geq 0,\ v(x,0)\geq 0,\ w(x,0)\geq 0, & x\in\Omega, \end{cases}$$

where $\Omega\subset\mathbb{R}^N$ is a bounded domain with smooth boundary $\partial\Omega$ and ν is the outward unit normal vector of the boundary $\partial\Omega$. The homogeneous Neumann boundary condition indicates that this system is self-contained with zero population flux across the boundary. The constants d_1, d_2, and d_3, called diffusion coefficients, are positive. By standard theory of parabolic equations we can prove that (1.5) has a unique classical solution (u, v, w) defined for all $t > 0$ and that this solution is positive when the initial data $u(x, 0)$, $v(x, 0)$, $w(x, 0)$ are nonnegative and are positive somewhere in Ω. It is obvious that $(\tilde{u}, \tilde{v}, \tilde{w})$ given by (1.4) is the only positive constant steady state of (1.5) if $m < abk$.

From the analysis in section 3, we can see that Turing patterns do not occur for (1.5), because the constant nonnegative steady states $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ have the same stability properties whether viewed as stationary solutions of (1.3) or (1.5). This suggests (though does not prove) that the dynamics of (1.5) could be similar to that of (1.3).

However, in (1.5), only diffusion of each individual species is taken into account. The reality of the interaction between the mature members and their young is that the latter tend to stay close to the former. We model this by the cross diffusion term $\Delta[d_4v/(\sigma + w^2)]$ for the immature predator, where d_4, called the cross diffusion coefficient, and σ are positive constants. Combined with (self-)diffusion, the immature predator thus diffuses with flux

$$J = -\nabla\left(d_2v + \frac{d_4v}{\sigma + w^2}\right) = -\left(d_2 + \frac{d_4}{\sigma + w^2}\right)\nabla v + \frac{2d_4vw}{(\sigma + w^2)^2}\nabla w.$$

We observe that, as $2d_4vw(\sigma+w^2)^{-2} \geq 0$, the part $\{2d_4vw(\sigma+w^2)^{-2}\}\nabla w$ of the flux is directed toward increasing w, that is, increasing population density of the mature predator.

Mathematically, this choice of the cross diffusion is one of the simplest functions that is biologically sound. There is also a technical point for this choice. We can replace w^2 in the cross diffusion term by w^τ, and the results in this paper can be proved for all $\tau > 1$. However, if $\tau \leq 1$, then our Propositions 2 and 3 in section 4 and Theorems 5 and 6 in section 7 are no longer true. Please also see section 9 for further discussion on the choice of the cross diffusion term.

Thus, the cross diffusion system that we shall study is the following:

$$\begin{cases} u_t - d_1\Delta u = G_1(\mathbf{u}), & x \in \Omega,\ t > 0, \\ v_t - \Delta\Big(d_2 v + \dfrac{d_4 v}{\sigma + w^2}\Big) = G_2(\mathbf{u}), & x \in \Omega,\ t > 0, \\ w_t - d_3\Delta w = G_3(\mathbf{u}), & x \in \Omega,\ t > 0, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x \in \partial\Omega,\ t > 0, \\ w(x,0) \geq 0,\ v(x,0) \geq 0,\ w(x,0) \geq 0, & x \in \Omega, \end{cases} \tag{1.6}$$

where $\mathbf{u} = (u, v, w)^T$ and

$$\mathbf{G}(\mathbf{u}) = \begin{pmatrix} G_1(\mathbf{u}) \\ G_2(\mathbf{u}) \\ G_3(\mathbf{u}) \end{pmatrix} = \begin{pmatrix} u(a - u - \varepsilon v - w) \\ kuw - v \\ bv - mw \end{pmatrix}.$$

The general theory in [1] guarantees that (1.6) has a unique nonnegative local solution (u, v, w). In this paper, we will mainly consider the steady-state solutions of (1.6), as a complete analysis of the dynamics of the system seems out of reach at the moment.

In section 2, we will determine the stability of $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ as stationary solutions of (1.3) and show that a Hopf bifurcation occurs. In section 3, we prove that $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ have the same stability properties when regarded as stationary solutions of (1.5); therefore, no Turing pattern can be found for the reaction-diffusion system (1.5). In the remaining sections, we study the problem (1.6). First, in section 4, we analyze the linearized eigenvalue problem of (1.6) at $\tilde{\mathbf{u}}$ in order to calculate the fixed point index of $\tilde{\mathbf{u}}$, which is important for our later discussions on the existence of nonconstant positive steady states, i.e., stationary patterns, of (1.6). In section 5, we establish a priori upper and lower bounds for all possible positive steady states of (1.6). In section 6, we establish the nonexistence of nonconstant positive steady states of (1.6) when the cross diffusion coefficient $d_4 = 0$, that is, when no cross diffusion occurs in the model. In section 7, we study the existence of nonconstant positive steady states for suitable values of the parameters. This is done by using the Leray–Schauder degree theory and the results obtained in sections 4 and 5. In section 8, we briefly discuss the bifurcation of nonconstant positive steady states of (1.6). This is followed by concluding discussions in section 9.

It is perhaps worth emphasizing that our results in sections 6 and 7 show that in (1.6) stationary patterns can arise in suitable parameter ranges only if cross diffusion is included in the model (i.e., $d_4 > 0$). This is in sharp contrast to most existing studies for predator-prey models with cross diffusion, such as [6, 7, 8, 39], where stationary patterns already arise with the introduction of the diffusion term for each species.

The mathematical approach in this paper is similar in spirit to that of [35], where a two-predator one-prey system was studied, and it was shown that the system has a unique positive constant steady state, which is the global attractor for the corresponding ODE and reaction-diffusion systems, but becomes unstable in certain parameter ranges when a cross diffusion term is included, and nonconstant positive stationary solutions emerge. However, our problem (1.6) is much more complex. For example, its ODE version, namely, (1.3), already exhibits much richer dynamical behavior (e.g., a Hopf bifurcation occurs; see section 2 for details). As a result, our analysis here, especially in sections 5 and 6, is much more difficult and requires techniques beyond those of [35]. On the other hand, our results here reinforce the point demonstrated in [35]; namely, cross diffusion may have crucial impact on the dynamics of certain predator-prey models.

2. Stability of $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ and Hopf bifurcation for the ODE system (1.3).

We first consider the stability of $\mathbf{u}_a$.

THEOREM 1. *If $m > abk$, then $\mathbf{u}_a$ is the only nontrivial nonnegative stationary solution of* (1.3)*, and every nonnegative solution (u, v, w) of* (1.3) *with $u \not\equiv 0$ satisfies* $\lim_{t\to\infty}(u, v, w) = (a, 0, 0)$.

Proof. As $m > abk$, there exists a $\delta > 0$ such that $m > (a + \delta)bk$. It follows from the first equation of (1.3) that $\limsup_{t\to\infty} u(t) \le a$. Therefore, we can find $T > 0$ such that $u(t) \le a + \delta$ for all $t \ge T$.

Consider the nonnegative solution $(\hat{v}, \hat{w})$ of the following problem:

$$\begin{cases} \dfrac{\mathrm{d}\hat{v}}{\mathrm{d}t} = k(a+\delta)\hat{w} - \hat{v}, & t > T, \\ \dfrac{\mathrm{d}\hat{w}}{\mathrm{d}t} = b\hat{v} - m\hat{w}, & t > T, \\ \hat{v}(T) = v(T), \quad \hat{w}(T) = w(T). \end{cases}$$

Then $v(t) \le \hat{v}(t)$ and $w(t) \le \hat{w}(t)$ for all $t \ge T$. As $(a + \delta)bk < m$, it is obvious that $\lim_{t\to\infty}(\hat{v}(t), \hat{w}(t)) = (0, 0)$. Consequently, $\lim_{t\to\infty}(v(t), w(t)) = (0, 0)$.

Applying the first equation of (1.3) once again, we deduce that $\lim_{t\to\infty} u(t) = a$. □

Next we consider the case $m < abk$. In this case, (1.3) has two nontrivial nonnegative stationary solutions: $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$.

THEOREM 2. *Suppose $m < abk$. Then $\mathbf{u}_a$ is unstable. Moreover,*

(i) *if*

$$b \le \varepsilon\left(1 + \frac{m}{bk}\right), \tag{2.1}$$

then $\tilde{\mathbf{u}}$ is asymptotically stable;

(ii) *if* (2.1) *does not hold, then there exists a unique $a^* > m/(bk)$, determined by ε, b, k, and m, such that $\tilde{\mathbf{u}}$ is asymptotically stable for $a \in (m/(bk), a^*)$ and is unstable for $a > a^*$; moreover, a branch of periodic solutions bifurcates from $\tilde{\mathbf{u}}$ at $a = a^*$.*

Proof. Since $m < abk$, one easily checks that the linearization matrix of (1.3) at $\mathbf{u}_a$ has one positive and two negative eigenvalues; therefore, $\mathbf{u}_a$ is unstable.

We now consider the stability of $\tilde{\mathbf{u}}$. A straightforward calculation shows that the linearization matrix of the system (1.3) at $\tilde{\mathbf{u}}$ has the characteristic polynomial $g(\lambda) = \lambda^3 + A_1\lambda^2 + A_2\lambda + A_3$, where

$$\begin{cases} A_1 = 1 + m + \dfrac{m}{bk} > 0, \\ A_2 = \dfrac{m}{bk}\left(1 + m + \dfrac{\varepsilon(abk - m)}{b + m\varepsilon}\right) > 0, \\ A_3 = \dfrac{m(abk - m)}{bk} > 0. \end{cases} \tag{2.2}$$

Therefore,

$$\begin{aligned}A_1A_2 - A_3 &= \frac{m}{bk(b+m\varepsilon)}\Big\{(1+m)(b+m\varepsilon)\Big(1+m+\frac{m}{bk}\Big)\\ &\qquad - (abk-m)\Big[b-\varepsilon\Big(1+\frac{m}{bk}\Big)\Big]\Big\}\\ &\triangleq \frac{m}{bk(b+m\varepsilon)}f(\xi), \quad \text{where } \xi = abk-m.\end{aligned}$$

Clearly, if (2.1) holds, then $f(\xi) > 0$ for all $\xi \geq 0$; hence $A_1A_2 - A_3 > 0$. If (2.1) does not hold, then $f(0) > 0$ and $f(\xi) < 0$ for all large ξ. Thus the linear function $f(\xi)$ has a unique positive zero point ξ_0 where $f(\xi) > 0$ for $\xi \in [0, \xi_0)$ and $f(\xi) < 0$ for $\xi > \xi_0$. It follows that $A_1A_2 - A_3 > 0$ when $a \in (m/(bk), a^*)$ and $A_1A_2 - A_3 < 0$ if $a > a^*$, where $a^* = (m + \xi_0)/(bk)$.

We can now apply the Routh–Hurwitz criterion (see, e.g., [27, Appendix 2]) to conclude that $\tilde{\mathbf{u}}$ is linearly stable (and hence asymptotically stable) if (2.1) holds or $a \in (m/(bk), a^*)$; and is linearly unstable (and hence unstable) if (2.1) does not hold and $a > a^*$.

It remains to show that a Hopf bifurcation occurs at $a = a^*$ when (2.1) does not hold. Since A_1, A_2, A_3 are all positive, clearly the characteristic polynomial $g(\lambda)$ has no nonnegative root. Since $g(\lambda) \to -\infty$ as $\lambda \to -\infty$, it always has a negative root. We now have two possibilities:

(i) $g(\lambda)$ has three negative roots: λ_1, λ_2, λ_3;

(ii) $g(\lambda)$ has one negative root λ_1 and a pair of complex roots: $\lambda_2 = \alpha + \beta i$, $\lambda_3 = \alpha - \beta i$, $\beta \neq 0$.

When a is close to a^*, $\lambda_1\lambda_2\lambda_3 = -A_3 < -(1/2)m\xi_0/(bk) < 0$. Therefore, no eigenvalue is close to 0 (in the complex plane) when a is close to a^*, say, $|\lambda_i| \geq \sigma_0 > 0$, $i = 1, 2, 3$. We show next that this implies that case (i) cannot happen when a is close to a^*. Indeed, if (i) happens with a close to a^*, then $A_1A_2 - A_3 = -\big[\lambda_1^2(\lambda_2+\lambda_3) + \lambda_2^2(\lambda_1+\lambda_3) + \lambda_3^2(\lambda_1+\lambda_2) + 2\lambda_1\lambda_2\lambda_3\big] \geq 8\sigma_0^3$ for all a close to a^*, which is impossible since, by our choice of a^*, $A_1A_2 - A_3 = 0$ when $a = a^*$. Therefore, case (ii) must happen when a is close to a^*. It then follows that $A_1A_2 - A_3 = -2\alpha[(\alpha+\lambda_1)^2 + \beta^2]$. If we regard α, β, and λ_1 as functions of ξ, we can rewrite the above identity as

$$\frac{m}{bk(b+m\varepsilon)}f(\xi) = -2\alpha(\xi)\big[(\alpha(\xi)+\lambda_1(\xi))^2 + \beta^2(\xi)\big].$$

It follows that

$$\alpha(\xi_0) = 0, \quad \frac{m}{bk(b+m\varepsilon)}f'(\xi_0) = -2\alpha'(\xi_0)[\lambda_1^2(\xi_0) + \beta^2(\xi_0)].$$

Hence $\mathrm{d}\alpha/\mathrm{d}a\big|_{a=a^*} = \alpha'(\xi_0)bk \neq 0$, and the Hopf bifurcation theorem can be applied to conclude that (1.3) has a branch of periodic solutions bifurcating from $\tilde{\mathbf{u}}$ at $a = a^*$. □

Remark 1. Note that if $\varepsilon \geq 1$, then (2.1) always holds (since $d < 1$); hence $\tilde{\mathbf{u}}$ is stable and Hopf bifurcation never occurs. On the other hand, if $\varepsilon = 0$, then (2.1) never holds.

3. Stability of $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ for the PDE system without cross diffusion (1.5).

In this section, we show that $\mathbf{u}_a$ and $\tilde{\mathbf{u}}$ have the same stability properties as in section 2.

First, assume that $abk < m$, and let (u, v, w) be an arbitrary nonnegative solution of (1.5) with $u \not\equiv 0$. From the first equation in (1.5) we deduce that $\limsup_{t\to\infty} u(x, t) \le a$ uniformly in $x \in \bar{\Omega}$. Therefore, we can find $T > 0$ such that $u(x, t) \le a + \delta$ for all $t \ge T$, where $\delta > 0$ is small so that $(a + \delta)bk < m$.

Consider the nonnegative solution $(\hat{v}, \hat{w})$ of the problem

$$\begin{cases} \dfrac{\mathrm{d}\hat{v}}{\mathrm{d}t} = k(a+\delta)\hat{w} - \hat{v}, & t > T, \\ \dfrac{\mathrm{d}\hat{w}}{\mathrm{d}t} = b\hat{v} - m\hat{w}, & t > T, \\ \hat{v}(T) = \max\limits_{x\in\bar{\Omega}} v(x, T), & \hat{w}(T) = \max\limits_{x\in\bar{\Omega}} w(x, T). \end{cases}$$

Then, $v(x, t) \le \hat{v}(t)$ and $w(x, t) \le \hat{w}(t)$ for all $t \ge T$ and $x \in \bar{\Omega}$. As $(a+\delta)bk < m$, we have, as before, $\lim_{t\to\infty}(\hat{v}(t), \hat{w}(t)) = (0, 0)$, and thus $\lim_{t\to\infty}(v(x, t), w(x, t)) = (0, 0)$ uniformly in $x \in \bar{\Omega}$.

Applying the first equation of (1.5) once again, we deduce that $\lim_{t\to\infty} u(x, t) = a$. Therefore, $\mathbf{u}_a$ is globally attractive, as in section 2.

Next we assume that $m < abk$. By Theorem 2, we find that $\mathbf{u}_a$ is unstable. It remains to check the stability of $\tilde{\mathbf{u}}$.

Let $0 = \mu_1 < \mu_2 < \mu_3 < \cdots$ be the eigenvalues of the operator $-\Delta$ on Ω with the homogeneous Neumann boundary condition, and let $E(\mu_i)$ be the eigenspace corresponding to μ_i in $H^1(\Omega)$. Let $\mathbf{X} = [H^1(\Omega)]^3$, $\{\phi_{ij}\,;\, j = 1, \ldots, \dim E(\mu_i)\}$ be an orthonormal basis of $E(\mu_i)$, and let $\mathbf{X}_{ij} = \{\mathbf{c}\phi_{ij} : \mathbf{c} \in \mathbb{R}^3\}$. Then,

$$\mathbf{X} = \bigoplus_{i=1}^{\infty} \mathbf{X}_i \qquad \text{and} \qquad \mathbf{X}_i = \bigoplus_{j=1}^{\dim E(\mu_i)} \mathbf{X}_{ij}. \tag{3.1}$$

Theorem 3. *Suppose that $m < abk$. Then the constant positive steady state $\tilde{\mathbf{u}}$ of* (1.5) *is linearly stable, and hence asymptotically stable in the sense of* [17], *when* (2.1) *holds or $a \in (m/(bk), a^*)$; and is unstable when* (2.1) *does not hold and $a > a^*$.*

Proof. We note that when $\tilde{\mathbf{u}}$ is unstable for (1.3), it is also unstable for (1.5). Thus, we need only to prove the asymptotic stability of $\tilde{\mathbf{u}}$ when our stability assumptions hold.

Let $\mathcal{D}$ =diag(d_1, d_2, d_3) and $\mathcal{L} = \mathcal{D}\Delta + \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$. The linearization of (1.5) at $\tilde{\mathbf{u}}$ is $\mathbf{u}_t = \mathcal{L}\mathbf{u}$. For each $i \ge 1$, $\mathbf{X}_i$ is invariant under the operator $\mathcal{L}$, and λ is an eigenvalue of $\mathcal{L}$ if and only if it is an eigenvalue of the matrix $-\mu_i\mathcal{D} + \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$ for some $i \ge 1$, in which case there is an eigenvector in $\mathbf{X}_i$. The characteristic polynomial of $-\mu_i\mathcal{D} + \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$ is given by $\psi_i(\lambda) = \lambda^3 + B_1\lambda^2 + B_2\lambda + B_3$, where

$$\begin{aligned} B_1 &= \mu_i(d_1 + d_2 + d_3) + 1 + m + \tilde{u} > 0, \\ B_2 &= \mu_i^2(d_1d_2 + d_1d_3 + d_2d_3) + \mu_i[(1+m)d_1 + (m+\tilde{u})d_2 + (1+\tilde{u})d_3] \\ &\qquad + (1 + m + \varepsilon k\tilde{w})\tilde{u} > 0, \\ B_3 &= \mu_i^3 d_1d_2d_3 + \mu_i^2(md_1d_2 + d_1d_3 + \tilde{u}d_2d_3) + \mu_i(md_2 + d_3 + \varepsilon k\tilde{w}d_3)\tilde{u} \\ &\qquad + (b + m\varepsilon)k\tilde{u}\tilde{w} > 0. \end{aligned}$$

A direct calculation yields $B_1B_2 - B_3 = c_3\mu_i^3 + c_2\mu_i^2 + c_1\mu_i + A_1A_2 - A_3$, where A_1, A_2, and A_3 are given by (2.2), and c_1, c_2, c_3 are positive. Hence $B_1B_2 - B_3 \ge A_1A_2 - A_3$. From the proof of Theorem 2 we see that $A_1A_2 - A_3 > 0$ under our conditions for stability. So, $B_1B_2 - B_3 > 0$. It thus follows from the Routh–Hurwitz criterion that, for each $i \geq 1$, the three roots $\lambda_{i,1}$, $\lambda_{i,2}$, $\lambda_{i,3}$ of $\psi_i(\lambda) = 0$ all have negative real parts.

In the following we shall prove that there exists a positive constant δ such that

$$\text{(3.2)} \qquad \mathrm{Re}\{\lambda_{i,1}\}, \quad \mathrm{Re}\{\lambda_{i,2}\}, \quad \mathrm{Re}\{\lambda_{i,3}\} \leq -\delta \qquad \forall\, i \geq 1.$$

Consequently, the spectrum of $\mathcal{L}$, which consists of eigenvalues, lies in $\{\mathrm{Re}\,\lambda \leq -\delta\}$, and the asymptotical stability of $\tilde{\mathbf{u}}$ follows [17, Theorem 5.1.1].

To see (3.2), let $\lambda = \mu_i\xi$. Then $\psi_i(\lambda) = \mu_i^3\xi^3 + B_1\mu_i^2\xi^2 + B_2\mu_i\xi + B_3 \overset{\Delta}{=} \tilde{\psi}_i(\xi)$. Since $\mu_i \to \infty$ as $i \to \infty$, it follows that $\lim_{i\to\infty}\{\tilde{\psi}_i(\xi)/\mu_i^3\} = \xi^3 + (d_1 + d_2 + d_3)\xi^2 + (d_1d_2 + d_1d_3 + d_2d_3)\xi + d_1d_2d_3 \overset{\Delta}{=} \bar{\psi}(\xi)$. Clearly $\bar{\psi}(\xi) = 0$ has three negative roots: $-d_1$, $-d_2$, $-d_3$. By continuity, we see that there exists i_0 such that the three roots $\xi_{i,1}$, $\xi_{i,2}$, $\xi_{i,3}$ of $\tilde{\psi}_i(\xi) = 0$ satisfy $\mathrm{Re}\{\xi_{i,1}\}$, $\mathrm{Re}\{\xi_{i,2}\}$, $\mathrm{Re}\{\xi_{i,3}\} \leq -\bar{\delta}/2$ for all $i \geq i_0$, where $\bar{\delta} = \min\{d_1, d_2, d_3\}$. In turn, $\mathrm{Re}\{\lambda_{i,1}\}$, $\mathrm{Re}\{\lambda_{i,2}\}$, $\mathrm{Re}\{\lambda_{i,3}\} \leq -\mu_i\bar{\delta}/2 \leq -\bar{\delta}/2$ for all $i \geq i_0$.

Let $-\tilde{\delta} = \max_{1\leq i\leq i_0}\{\mathrm{Re}\{\lambda_{i,1}\}, \mathrm{Re}\{\lambda_{i,2}\}, \mathrm{Re}\{\lambda_{i,3}\}\}$. Then $\tilde{\delta} > 0$, and (3.2) holds for $\delta = \min\{\tilde{\delta},\, \bar{\delta}/2\}$. □

4. Fixed point index of ũ for the stationary PDE system with cross diffusion. Let $\Phi(\mathbf{u}) = (d_1u,\, d_2v + d_4v/(\sigma + w^2),\, d_3w)^T$. Then the stationary problem of (1.6) can be written as

$$\text{(4.1)} \qquad -\Delta\Phi(\mathbf{u}) = \mathbf{G}(\mathbf{u}), \quad x \in \Omega; \qquad \frac{\partial \mathbf{u}}{\partial \nu} = 0, \quad x \in \partial\Omega.$$

In this section, we study the linearization of (4.1) at $\tilde{\mathbf{u}}$ and then proceed to calculate the fixed point index of $\tilde{\mathbf{u}}$ when it is an isolated solution.

Let $\mathbf{Y} = [C^1(\bar{\Omega})]^3$, and define $\mathbf{Y}^+ = \{\mathbf{u} \in \mathbf{Y} : u, v, w > 0 \text{ on } \bar{\Omega}\}$ and, for $C > 0$, $B(C) = \{\mathbf{u} \in \mathbf{Y} : C^{-1} < u, v, w < C \text{ on } \bar{\Omega}\}$.

Since the determinant of $\Phi_{\mathbf{u}}(\mathbf{u})$ is positive for all nonnegative $\mathbf{u}$, $\Phi_{\mathbf{u}}^{-1}(\mathbf{u})$ exists and $\det \Phi_{\mathbf{u}}^{-1}(\mathbf{u})$ is positive. Hence, $\mathbf{u}$ is a positive solution to (4.1) if and only if

$$\mathbf{F}(\mathbf{u}) \overset{\Delta}{=} \mathbf{u} - (\mathbf{I} - \Delta)^{-1}\{\Phi_{\mathbf{u}}^{-1}(\mathbf{u})[\mathbf{G}(\mathbf{u}) + \nabla\mathbf{u}\,\Phi_{\mathbf{uu}}(\mathbf{u})\nabla\mathbf{u}] + \mathbf{u}\} = 0 \ \text{ in } \ \mathbf{Y}^+,$$

where $(\mathbf{I} - \Delta)^{-1}$ is the inverse of $\mathbf{I} - \Delta$ under homogeneous Neumann boundary conditions. As $\mathbf{F}(\cdot)$ is a compact perturbation of the identity operator, for any $B = B(C)$, the Leray–Schauder degree $\deg(\mathbf{F}(\cdot),\, 0,\, B)$ is well defined if $\mathbf{F}(\mathbf{u}) \neq 0$ on ∂B.

Further, we note that $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\}$ and recall that, if $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ is invertible, the fixed point index of $\mathbf{F}$ at $\tilde{\mathbf{u}}$ is well defined and

$$\mathrm{index}(\mathbf{F}(\cdot),\, \tilde{\mathbf{u}}) = (-1)^{\gamma},$$

where γ is the sum of the algebraic multiplicities of all the negative eigenvalues of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ [29, Theorem 2.8.1].

Since the eigenvalues of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ and their algebraic multiplicities are the same regardless of whether it is considered an operator in $\mathbf{X}$ or in $\mathbf{Y}$, it is convenient to use the decomposition (3.1) in our discussion of the eigenvalues of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$. A straightforward calculation shows that, for each integer $i \geq 1$ and each integer $1 \leq j \leq \dim E(\mu_i)$, $\mathbf{X}_{ij}$ is invariant under $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$. Moreover, λ is an eigenvalue of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ if and only if, for some $i \geq 1$, it is an eigenvalue of the matrix

$$\mathbf{B}_i := \mathbf{I} - \frac{1}{1 + \mu_i}[\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}] = \frac{1}{1 + \mu_i}[\mu_i\mathbf{I} - \Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})].$$

Thus, $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ is invertible if and only if the matrix $\mathbf{B}_i$ is nonsingular for all $i \geq 1$.

Let λ be an eigenvalue of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$. We now calculate its algebraic multiplicity, which we denote by $\sigma(\lambda)$. By definition, $\sigma(\lambda) = \dim(E^{\lambda})$, where $E^{\lambda} := \bigcup_{n=1}^{\infty} \operatorname{Ker}\big[\lambda I - D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})\big]^n$.

For any $\phi \in \mathbf{X}$, we can uniquely express it in the form $\phi = \sum_{i=1}^{\infty}\sum_{j=1}^{\dim E(\mu_i)} C_{ij}\phi_{ij}$, where $C_{ij} \in \mathbb{R}^3$ and ϕ_{ij} is defined as before. Since $\mathbf{X}_{ij}$ is invariant under $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$, it is also invariant under $\big[\lambda I - D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})\big]^n$ for any $n \geq 1$. Thus, $\phi \in \operatorname{Ker}\big[\lambda I - D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})\big]^n$ if and only if $\big[\lambda I - D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})\big]^n C_{ij}\phi_{ij} = 0$ for all $i \geq 1$ and $1 \leq j \leq \dim E(\mu_i)$. A direct calculation shows that $\big[\lambda I - D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})\big]^n C_{ij}\phi_{ij} = 0$ if and only if $\big[\lambda I - \mathbf{B}_i\big]^n C_{ij} = 0$. It follows that

$$\dim E^{\lambda} = \sum_{i=1}^{\infty}\left[\dim E(\mu_i) \times \dim\left(\bigcup_{n=1}^{\infty}\operatorname{Ker}(\lambda\mathbf{I} - \mathbf{B}_i)^n\right)\right].$$

Now, $\dim\big(\bigcup_{n=1}^{\infty}\operatorname{Ker}(\lambda I - \mathbf{B}_i)^n\big)$ is just the algebraic multiplicity of λ as an eigenvalue of the 3×3 matrix $\mathbf{B}_i$. Writing

$$H(\mu) = H(\tilde{\mathbf{u}};\, \mu) \overset{\Delta}{=} \det\big\{\mu\,\mathbf{I} - \Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\big\} = \det\big\{(\mu - \mu_i)\mathbf{I} + (1+\mu_i)\mathbf{B}_i\big\}, \tag{4.2}$$

we easily see that $\lambda = (\mu_i - \mu)/(1+\mu_i)$ is an eigenvalue of $\mathbf{B}_i$ if and only if $H(\mu) = 0$. Moreover, if $H(\mu_i) \neq 0$, then the number of negative eigenvalues (counting algebraic multiplicity) of $\mathbf{B}_i$ is odd if and only if $H(\mu_i) < 0$. Therefore,

$$\sigma(\lambda) = \dim E^{\lambda} = \sum_{i\geq 1,\, H(\mu_i)<0} \dim E(\mu_i) \pmod 2.$$

As a consequence, we have the following proposition.

Proposition 1. *Suppose that, for all $i \geq 1$, $H(\mu_i) \neq 0$. Then*

$$\mathit{index}(\mathbf{F}(\cdot),\, \tilde{\mathbf{u}}) = (-1)^{\gamma}, \qquad \text{where} \qquad \gamma = \sum_{i\geq 1,\, H(\mu_i)<0} \dim E(\mu_i).$$

To facilitate our computation of index$(\mathbf{F}(\cdot),\, \tilde{\mathbf{u}})$, we need to determine the sign of $H(\mu_i)$. In particular, as the aim of this paper is to study the existence of stationary patterns of (1.6) with respect to the cross diffusion coefficient d_4 and diffusion coefficient d_1, we will concentrate on the dependence of $H(\mu_i)$ on d_4 and d_1. At this point, we note that $H(\mu) = \det\{\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\}\ \det\{\mu\,\Phi_{\mathbf{u}}(\tilde{\mathbf{u}}) - \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\}$. Since we have already established that $\det\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})$ is positive, we will need only to consider $\det\{\mu\,\Phi_{\mathbf{u}}(\tilde{\mathbf{u}}) - \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\}$.

As

$$\Phi_{\mathbf{u}}(\tilde{\mathbf{u}}) = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 + \dfrac{d_4}{\sigma + \tilde{w}^2} & -\dfrac{2d_4\tilde{v}\tilde{w}}{(\sigma+\tilde{w}^2)^2} \\ 0 & 0 & d_3 \end{pmatrix},$$

we have

$$\begin{aligned} \det\{\mu\Phi_{\mathbf{u}}(\tilde{\mathbf{u}}) - \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} &= C_3(\sigma, d_1, d_4)\mu^3 + C_2(\sigma, d_1, d_4)\mu^2 + C_1(\sigma, d_1, d_4)\mu \\ &\quad + k(b + m\varepsilon)\tilde{u}\tilde{w} \\ &\overset{\Delta}{=} \mathcal{C}(\sigma, d_1, d_4;\, \mu), \end{aligned} \tag{4.3}$$

where

$$\begin{aligned}
C_3(\sigma,d_1,d_4) &= d_1d_3\Big(d_2+\frac{d_4}{\sigma+\tilde{w}^2}\Big),\\
C_2(\sigma,d_1,d_4) &= (md_1+d_3\tilde{u})\Big(d_2+\frac{d_4}{\sigma+\tilde{w}^2}\Big)+d_1d_3-bd_1\frac{2d_4\tilde{v}\tilde{w}}{(\sigma+\tilde{w}^2)^2},\\
C_1(\sigma,d_1,d_4) &= m\tilde{u}\Big(d_2+\frac{d_4}{\sigma+\tilde{w}^2}\Big)+d_3\tilde{u}-b\tilde{u}\frac{2d_4\tilde{v}\tilde{w}}{(\sigma+\tilde{w}^2)^2}+\varepsilon kd_3\tilde{u}\tilde{w}.
\end{aligned}$$

First we consider the dependence of $\mathcal{C}$ on d_4. Let $\tilde{\mu}_1(d_4)$, $\tilde{\mu}_2(d_4)$, and $\tilde{\mu}_3(d_4)$ be the three roots of $\mathcal{C}(\sigma,d_1,d_4;\,\mu)=0$ with $\mathrm{Re}\{\tilde{\mu}_1(d_4)\}\le\mathrm{Re}\{\tilde{\mu}_2(d_4)\}\le\mathrm{Re}\{\tilde{\mu}_3(d_4)\}$. Then $\tilde{\mu}_1(d_4)\tilde{\mu}_2(d_4)\tilde{\mu}_3(d_4)<0$, at least one of $\tilde{\mu}_1(d_4)$, $\tilde{\mu}_2(d_4)$, $\tilde{\mu}_3(d_4)$ is real and negative, and the product of the other two is positive.

Consider the following limits:

$$\begin{aligned}
\lim_{d_4\to\infty}\frac{C_3(\sigma,d_1,d_4)}{d_4} &= \frac{d_1d_3}{\sigma+\tilde{w}^2}\ \triangleq\ a_3(\sigma,d_1),\\
\lim_{d_4\to\infty}\frac{C_2(\sigma,d_1,d_4)}{d_4} &= \frac{1}{\sigma+\tilde{w}^2}\Big(md_1+d_3\tilde{u}-2bd_1\frac{\tilde{v}\tilde{w}}{\sigma+\tilde{w}^2}\Big)\triangleq\ a_2(\sigma,d_1),\\
\lim_{d_4\to\infty}\frac{C_1(\sigma,d_1,d_4)}{d_4} &= \frac{\tilde{u}[m(\sigma+\tilde{w}^2)-2b\tilde{v}\tilde{w}]}{(\sigma+\tilde{w}^2)^2}=\frac{m\tilde{u}(\sigma-\tilde{w}^2)}{(\sigma+\tilde{w}^2)^2}\triangleq\ a_1(\sigma).
\end{aligned}$$

Therefore, $a_1(\sigma)<0$ when $\sigma<\tilde{w}^2$. In the following, we restrict our attention to $0<\sigma<\tilde{w}^2$. In this range, $a_1(\sigma)<0$, and $C_1(\sigma,d_1,d_4)<0$ for all sufficiently large d_4. Note that

$$\begin{aligned}
\lim_{d_4\to\infty}\frac{\mathcal{C}(\sigma,d_1,d_4;\,\mu)}{d_4} &= a_3(\sigma,d_1)\mu^3+a_2(\sigma,d_1)\mu^2+a_1(\sigma)\mu\\
&= \mu[a_3(\sigma,d_1)\mu^2+a_2(\sigma,d_1)\mu+a_1(\sigma)],
\end{aligned}$$

and $a_1(\sigma)<0<a_3(\sigma,d_1)$. A continuity argument shows that, when d_4 is large, $\tilde{\mu}_1(d_4)$ is real and negative. Furthermore, as $\tilde{\mu}_2(d_4)\tilde{\mu}_3(d_4)>0$, $\tilde{\mu}_2(d_4)$ and $\tilde{\mu}_3(d_4)$ are real and positive, and

$$(4.4)\qquad \begin{cases}
\displaystyle\lim_{d_4\to\infty}\tilde{\mu}_1(d_4)=\frac{-a_2(\sigma,d_1)-\sqrt{a_2^2(\sigma,d_1)-4a_1(\sigma)a_3(\sigma,d_1)}}{2a_3(\sigma,d_1)}<0,\\
\displaystyle\lim_{d_4\to\infty}\tilde{\mu}_2(d_4)=0,\\
\displaystyle\lim_{d_4\to\infty}\tilde{\mu}_3(d_4)=\frac{-a_2(\sigma,d_1)+\sqrt{a_2^2(\sigma,d_1)-4a_1(\sigma)a_3(\sigma,d_1)}}{2a_3(\sigma,d_1)}\triangleq\tilde{\mu}>0.
\end{cases}$$

Thus we have the following proposition.

PROPOSITION 2. *Assume that* $0<\sigma<\tilde{w}^2$. *Then there exists a positive number* d_4^* *such that, when* $d_4\ge d_4^*$, *the three roots* $\tilde{\mu}_1(d_4)$, $\tilde{\mu}_2(d_4)$, $\tilde{\mu}_3(d_4)$ *of* $\mathcal{C}(\sigma,d_1,d_4;\,\mu)=0$ *are all real and satisfy* (4.4). *Moreover, for all* $d_4\ge d_4^*$,

$$(4.5)\qquad \begin{cases}
-\infty<\tilde{\mu}_1(d_4)<0<\tilde{\mu}_2(d_4)<\tilde{\mu}_3(d_4),\\
\mathcal{C}(\sigma,d_1,d_4;\,\mu)<0\quad\textit{when}\quad \mu\in(-\infty,\,\tilde{\mu}_1(d_4))\cup(\tilde{\mu}_2(d_4),\,\tilde{\mu}_3(d_4)),\\
\mathcal{C}(\sigma,d_1,d_4;\,\mu)>0\quad\textit{when}\quad \mu\in(\tilde{\mu}_1(d_4),\,\tilde{\mu}_2(d_4))\cup(\tilde{\mu}_3(d_4),\,\infty).
\end{cases}$$

Next we consider the dependence of $\mathcal{C}$ on d_1. In this case, we consider the limits

$$\begin{aligned}
\lim_{d_1\to\infty} \frac{C_3(\sigma,d_1,d_4)}{d_1} &= d_3\Big(d_2+\frac{d_4}{\sigma+\tilde{w}^2}\Big) \triangleq b_3(\sigma,d_4),\\
\lim_{d_1\to\infty} \frac{C_2(\sigma,d_1,d_4)}{d_1} &= md_2+d_3+\frac{md_4}{(\sigma+\tilde{w}^2)^2}(\sigma-\tilde{w}^2) \triangleq b_2(\sigma,d_4),\\
\lim_{d_1\to\infty} \frac{C_1(\sigma,d_1,d_4)}{d_1} &= 0,\\
\lim_{d_1\to\infty} \frac{\mathcal{C}(\sigma,d_1,d_4;\,\mu)}{d_1} &= \mu^2[b_3(\sigma,d_4)\mu+b_2(\sigma,d_4)].
\end{aligned}$$

When the parameters satisfy $b_2(\sigma,d_4)<0$, i.e.,

$$md_2+d_3+\frac{m\sigma d_4}{(\sigma+\tilde{w}^2)^2} < \frac{m\tilde{w}^2 d_4}{(\sigma+\tilde{w}^2)^2}, \tag{4.6}$$

one can establish the following similarly to Proposition 2.

Proposition 3. *Assume that* (4.6) *holds. Then there exists a positive constant d_1^* such that, when $d_1\ge d_1^*$, the three roots $\bar\mu_1(d_1)$, $\bar\mu_2(d_1)$, $\bar\mu_3(d_1)$ of $\mathcal{C}(\sigma,d_1,d_4;\,\mu)=0$ are all real and satisfy* $\lim_{d_1\to\infty}\bar\mu_1(d_1)=\lim_{d_1\to\infty}\bar\mu_2(d_1)=0$ *and*

$$\lim_{d_1\to\infty}\bar\mu_3(d_1)=\frac{-b_2(\sigma,d_4)}{b_3(\sigma,d_4)} \triangleq \bar\mu>0. \tag{4.7}$$

Moreover, when $d_1\ge d_1^*$, we have

$$\begin{cases} -\infty<\bar\mu_1(d_1)<0<\bar\mu_2(d_1)<\bar\mu_3(d_1),\\ \mathcal{C}(\sigma,d_1,d_4;\,\mu)<0 \quad when \quad \mu\in(-\infty,\,\bar\mu_1(d_1))\cup(\bar\mu_2(d_1),\,\bar\mu_3(d_1)),\\ \mathcal{C}(\sigma,d_1,d_4;\,\mu)>0 \quad when \quad \mu\in(\bar\mu_1(d_1),\,\bar\mu_2(d_1))\cup(\bar\mu_3(d_1),\,\infty). \end{cases} \tag{4.8}$$

Remark 2. Assume that $\sigma<\tilde{w}^2$. If (i) d_4 is large, or (ii) d_4 is positive and d_2,d_3 are sufficiently small, then the inequality (4.6) holds.

5. A priori estimates. The main purpose of this section is to give a priori positive upper and lower bounds for the positive solutions to (4.1). For this, we shall make use of the following two results.

Proposition 4 (Harnack inequality [21]). *Let $z\in C^2(\Omega)\cap C^1(\bar\Omega)$ be a positive solution to $\Delta z(x)+c(x)z(x)=0$, where $c\in C(\bar\Omega)$, satisfying the homogeneous Neumann boundary condition. Then there exists a positive constant C which depends only on B where $\|c\|_\infty\le B$ such that* $\max_{\bar\Omega} z\le C\min_{\bar\Omega} z$.

Proposition 5 (maximum principle [25]). *Suppose that $g\in C(\bar\Omega)$ and $b_j\in C(\bar\Omega)$, $j=1,2,\dots,N$.*

(i) *If $z\in C^2(\Omega)\cap C^1(\bar\Omega)$ satisfies*

$$\Delta z(x)+\sum_{j=1}^N b_j(x)z_{x_j}+g(x)\ge 0 \quad in \quad \Omega, \qquad \frac{\partial z}{\partial\nu}\Big|_{\partial\Omega}\le 0,$$

and $z(x_0)=\max_{\bar\Omega} z$, then $g(x_0)\ge 0$.

(ii) *If $z \in C^2(\Omega) \cap C^1(\bar{\Omega})$ satisfies*

$$\Delta z(x) + \sum_{j=1}^{N} b_j(x) z_{x_j} + g(x) \le 0 \quad in \ \ \Omega, \qquad \frac{\partial z}{\partial \nu}\Big|_{\partial\Omega} \ge 0,$$

and $z(x_0) = \min_{\bar{\Omega}} z$, then $g(x_0) \le 0$.

For clarity, we write the problem (4.1) explicitly:

$$(5.1) \qquad \begin{cases} -d_1 \Delta u = au - u^2 - \varepsilon uv - uw, & x \in \Omega, \\ -\Delta \left(d_2 v + \dfrac{d_4 v}{\sigma + w^2} \right) = kuw - v, & x \in \Omega, \\ -d_3 \Delta w = bv - mw, & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x \in \partial\Omega. \end{cases}$$

In the rest of this section, we fix the parameters a, b, k, m, ε, and σ and estimate the upper and lower positive bounds of positive solutions to (5.1) with respect to the diffusion and cross diffusion coefficients d_i. The generic constants C, C_i to be used below will depend on the parameters a, b, k, m, ε, and σ. As they are fixed, this dependence will not be stated explicitly.

Proposition 6. *Let d and d^* be two fixed positive constants. Then there is a positive constant $C(d, d^*)$ such that, for any $d_1, d_2, d_3 \ge d$, and $0 \le d_4 \le d^*$, every possible positive solution (u, v, w) of* (5.1) *satisfies*

$$(5.2) \qquad \|u, v, w\|_{C^{2+\alpha}(\bar{\Omega})} \le C(d, d^*).$$

Proof. We first prove the estimate

$$(5.3) \qquad \max_{\bar{\Omega}} u, \ \max_{\bar{\Omega}} v, \ \max_{\bar{\Omega}} w \le C(d, d^*).$$

Applying the maximum principle to the equation of u, it is directly deduced that $\max_{\bar{\Omega}} u \le a$. If the estimate (5.3) is not true, then there exist $(d_{1n}, d_{2n}, d_{3n}, d_{4n})$ satisfying d_{1n}, d_{2n}, $d_{3n} \ge d$, and $0 \le d_{4n} \le d^*$, and a corresponding positive solution (u_n, v_n, w_n) of (5.1) with $(d_1, d_2, d_3, d_4) = (d_{1n}, d_{2n}, d_{3n}, d_{4n})$, such that

$$(5.4) \qquad \max_{\bar{\Omega}} v_n + \max_{\bar{\Omega}} w_n \to \infty \quad \text{as } n \to \infty.$$

Applying the maximum principle to the equation of w_n, we have

$$(5.5) \qquad \max_{\bar{\Omega}} w_n \le (b/m) \max_{\bar{\Omega}} v_n.$$

Let $\varphi_n = d_{2n} v_n + \frac{d_{4n} v_n}{\sigma + w_n^2}$ and $x_0 \in \bar{\Omega}$ be such that $\varphi_n(x_0) = \max_{\bar{\Omega}} \varphi_n$. Applying the maximum principle to the equation of v_n, we have $v_n(x_0) \le k u_n(x_0) w_n(x_0) \le ka \max_{\bar{\Omega}} w_n$. Hence,

$$\begin{aligned} d_{2n} \max_{\bar{\Omega}} v_n \le \max_{\bar{\Omega}} \varphi_n = \varphi_n(x_0) &= d_{2n} v_n(x_0) + \frac{d_{4n} v_n(x_0)}{\sigma + w_n^2(x_0)} \\ &\le v_n(x_0) \left(d_{2n} + \frac{d^*}{\sigma} \right) \le ka \max_{\bar{\Omega}} w_n \left(d_{2n} + \frac{d^*}{\sigma} \right), \end{aligned}$$

which implies that

$$\max_{\bar\Omega} v_n \le ka\left(1+\frac{d^*}{\sigma d}\right)\max_{\bar\Omega} w_n. \tag{5.6}$$

It follows from (5.4)–(5.6) that $\lim_{n\to\infty}\max_{\bar\Omega} v_n = \lim_{n\to\infty}\max_{\bar\Omega} w_n = \infty$. Set $\hat v_n = \frac{v_n}{\|v_n\|_\infty}$ and $\hat w_n = \frac{w_n}{\|w_n\|_\infty}$. Then $(u_n, \hat v_n, \hat w_n)$ satisfies

$$\begin{cases} -d_{1n}\Delta u_n = u_n\big[a-u_n-\varepsilon\|v_n\|_\infty \hat v_n - \|w_n\|_\infty \hat w_n\big], & x\in\Omega,\\ -\Delta\left(\hat v_n + \dfrac{d_{4n}\hat v_n}{d_{2n}(\sigma+\|w_n\|_\infty^2\hat w_n^2)}\right) = \dfrac{1}{d_{2n}}\left(ku_n\dfrac{\|w_n\|_\infty}{\|v_n\|_\infty}\hat w_n - \hat v_n\right), & x\in\Omega,\\ -\Delta \hat w_n = \dfrac{1}{d_{3n}}\left(b\dfrac{\|v_n\|_\infty}{\|w_n\|_\infty}\hat v_n - m\hat w_n\right), & x\in\Omega,\\ \|\hat v_n\|_\infty = \|\hat w_n\|_\infty = 1, \quad \dfrac{\partial u_n}{\partial\nu} = \dfrac{\partial \hat v_n}{\partial\nu} = \dfrac{\partial \hat w_n}{\partial\nu} = 0, & x\in\partial\Omega. \end{cases} \tag{5.7}$$

In view of (5.5) and (5.6) we have $A\|v_n\|_\infty \le \|w_n\|_\infty \le B\|v_n\|_\infty$ for some positive constants A and B. Since $0\le u_n\le a$ and $0\le \hat v_n, \hat w_n \le 1$, subject to a subsequence we may assume that $\|v_n\|_\infty/\|w_n\|_\infty \to \theta$ for some positive constant θ, and

$$\begin{gathered} d_{in}\to d_i \quad \text{with } d_1,d_2,d_3\ge d, \text{ and } 0\le d_4\le d^*,\\ u_n\to u \ \text{ strongly in } \ L^p(\Omega), \quad \hat v_n\to\hat v \ \text{ weakly in } \ L^p(\Omega),\\ \hat\varphi_n\to\hat\varphi,\quad \hat w_n\to\hat w \quad \text{weakly in } \ W^{2,p}(\Omega), \quad \text{and } \ \|\hat w\|_\infty = 1, \end{gathered}$$

where $p>N$ and $\hat\varphi_n = \hat v_n + \frac{d_{4n}\hat v_n}{d_{2n}(\sigma+\|w_n\|_\infty^2\hat w_n^2)}$. Hence $\hat\varphi$, $\hat w\in C^{1+\alpha}(\bar\Omega)$ for some $\alpha>0$, and $\hat\varphi_n\to\hat\varphi$, $\hat w_n\to\hat w$ in $C^{1+\alpha}(\bar\Omega)$.

If $d_3=\infty$, then $\hat w$ satisfies

$$-\Delta\hat w = 0 \quad \text{in } \ \Omega, \qquad \frac{\partial\hat w}{\partial\nu} = 0 \quad \text{on } \ \partial\Omega, \qquad \|\hat w\|_\infty = 1,$$

which implies that $\hat w = 1$. If $d_3<\infty$, then $\hat w$ satisfies $\|\hat w\|_\infty = 1$ and

$$-d_3\Delta\hat w = b\theta\hat v - m\hat w \quad \text{in } \ \Omega, \qquad \frac{\partial\hat w}{\partial\nu} = 0 \quad \text{on } \ \partial\Omega. \tag{5.8}$$

By the strong maximum principle and the Hopf boundary lemma for the $W^{2,N}(\Omega)$ solution (see [15, Theorem 9.6] and [12, Theorem 2.11]), we see that $\hat w>0$ on $\bar\Omega$.

The above analysis shows that, for both cases $d_3=\infty$ and $d_3<\infty$, we have $\hat w>0$ on $\bar\Omega$. Hence, there is a positive constant δ such that $\hat w\ge\delta$ on $\bar\Omega$. Consequently, $\hat w_n\ge\delta/2$ on $\bar\Omega$ for all large n. Since $\|w_n\|_\infty\to\infty$ as $n\to\infty$, from the equation of u_n in (5.7) we have that, for large n,

$$\begin{cases} -d_{1n}\Delta u_n = u_n\big[a-u_n-\varepsilon\|v_n\|_\infty\hat v_n - \|w_n\|_\infty\hat w_n\big]\\ \qquad\qquad\quad \le u_n\big[a-(\delta/2)\|w_n\|_\infty\big] < 0, \qquad x\in\Omega,\\ \dfrac{\partial u_n}{\partial\nu} = 0, \qquad x\in\partial\Omega. \end{cases}$$

This is impossible since $\int_\Omega \Delta u_n dx = 0$.

Now, we prove the estimate (5.2). Due to (5.3), by the regularity for elliptic equations we have that u, w, and $v\left[d_2 + d_4/(\sigma + w^2)\right]$ belong to $C^{1+\alpha}(\bar{\Omega})$, and the $C^{1+\alpha}(\bar{\Omega})$ norms of them depend only on the parameters d, d^* and the parameters $a, b, k, m, \varepsilon, \sigma$. It follows that $v \in C^{1+\alpha}(\bar{\Omega})$ and $\|v\|_{C^{1+\alpha}(\bar{\Omega})}$ depends only on the parameters d, d^* and $a, b, k, m, \varepsilon, \sigma$. Using the regularity of elliptic equations again, the estimate (5.2) follows. □

In the following we estimate the positive lower bound of positive solutions. We first state a lemma whose proof we shall omit.

Lemma 1. *Let $d_{in} \in (0, \infty)$, $i = 1, 2, 3, 4$, and (u_n, v_n, w_n) be the corresponding positive solution of* (5.1) *with $d_i = d_{in}$. Assume that $d_{in} \to d_i \in [0, \infty]$ and $(u_n, v_n, w_n) \to (u^*, v^*, w^*)$ uniformly on $\bar{\Omega}$. If u^*, v^*, and w^* are constants, then (u^*, v^*, w^*) must satisfy $a - u^* - \varepsilon v^* - w^* = 0$, $ku^*w^* - v^* = 0$, and $bv^* - mw^* = 0$. In particular, if u^*, v^*, and w^* are positive constants, then $(u^*, v^*, w^*) = (\tilde{u}, \tilde{v}, \tilde{w})$, the unique positive constant solution of* (5.1).

Proposition 7. *Let d and d^* be two fixed positive constants. Then there is a positive constant $C(d, d^*)$ such that, for any $d_1, d_2, d_3 \geq d$, and $0 \leq d_4 \leq d^*$, every possible positive solution (u, v, w) of* (5.1) *satisfies*

$$\min_{\bar{\Omega}} u,\ \min_{\bar{\Omega}} v,\ \min_{\bar{\Omega}} w \geq \frac{1}{C(d, d^*)}.$$

Proof. If the conclusion does not hold, then there exist a sequence $\{(d_{1n}, d_{2n}, d_{3n}, d_{4n})\}$ which satisfies $d_{1n}, d_{2n}, d_{3n} \geq d$, and $0 \leq d_{4n} \leq d^*$ and a sequence of corresponding positive solutions (u_n, v_n, w_n) of (5.1) with $d_i = d_{in}$, such that $\min\{\min_{\bar{\Omega}} u_n, \min_{\bar{\Omega}} v_n, \min_{\bar{\Omega}} w_n\} \to 0$. As $d_{1n}, d_{2n}, d_{3n} \geq d$, subject to a subsequence, we may assume that $d_{in} \to d_i \in [d, \infty]$ for $i = 1, 2, 3$, and $d_{4n} \to d_4 \in [0, d^*]$. By (5.2), we may also assume that $(u_n, v_n, w_n) \to (u, v, w)$ in $[C^{2+\alpha}(\bar{\Omega})]^3$ for some nonnegative functions u, v, w. It is easy to see that (u, v, w) also satisfies the estimate (5.2), and $\min_{\bar{\Omega}} u = 0$ or $\min_{\bar{\Omega}} v = 0$ or $\min_{\bar{\Omega}} w = 0$. Moreover, we observe that, if $d_1, d_2, d_3 < \infty$, then (u, v, w) satisfies (5.1). If $d_1 = \infty$, as (u_n, v_n, w_n) satisfies (5.3), then u satisfies $-\Delta u = 0$ in Ω and $\partial u/\partial \nu = 0$ on $\bar{\Omega}$, and hence u is constant. Analogous conclusions hold for d_2 and d_3.

Next we derive a contradiction for every possible case.

Step 1. We consider the case $d_1, d_2, d_3 < \infty$. First, in view of (5.3), the Harnack inequality shows $\min_{\bar{\Omega}} u = 0$ implies that $u = 0$ on $\bar{\Omega}$. In that case, by the strong maximum principle and the Hopf boundary lemma, it follows that $v = w = 0$ on $\bar{\Omega}$. But this is a contradiction to Lemma 1. Thus, $\min_{\bar{\Omega}} u > 0$.

Next, we show that $\min_{\bar{\Omega}} v = \min_{\bar{\Omega}} w = 0$. By our assumption, at least one of these is 0. If $\min_{\bar{\Omega}} w = 0$, we denote $w(x_0) = \min_{\bar{\Omega}} w$. By the maximum principle we have $bv(x_0) \leq mw(x_0) = 0$, and so $\min_{\bar{\Omega}} v = 0$. Conversely, if $v(x_1) = \min_{\bar{\Omega}} v = 0$, then $v(x_1)[d_2 + d_4/(\sigma + w^2(x_1))] = 0 = \min_{\bar{\Omega}} v[d_2 + d_4/(\sigma + w^2)]$. Applying the maximum principle to the second equation of (5.1), we have $ku(x_1)w(x_1) \leq v(x_1) = 0$. As $u(x_1) > 0$, we conclude that $w(x_1) = 0$, and so $\min_{\bar{\Omega}} w = 0$. In conclusion, we always have $\min_{\bar{\Omega}} v = \min_{\bar{\Omega}} w = 0$; that is, $\lim_{n\to\infty} \min_{\bar{\Omega}} v_n = \lim_{n\to\infty} \min_{\bar{\Omega}} w_n = 0$.

Define

$$\hat{v}_n = \frac{v_n}{\|v_n\|_\infty + \|w_n\|_\infty}, \quad \hat{w}_n = \frac{w_n}{\|v_n\|_\infty + \|w_n\|_\infty}. \tag{5.9}$$

Then $(u_n, \hat{v}_n, \hat{w}_n, w_n)$ satisfies

$$\begin{cases} -d_{1n}\Delta u_n = u_n(a - u_n - \varepsilon v_n - w_n), & x \in \Omega, \\ -\Delta\left(d_{2n}\hat{v}_n + \dfrac{d_{4n}\hat{v}_n}{\sigma + w_n^2}\right) = ku_n\hat{w}_n - \hat{v}_n, & x \in \Omega, \\ -d_{3n}\Delta\hat{w}_n = b\hat{v}_n - m\hat{w}_n, & x \in \Omega, \\ \dfrac{\partial u_n}{\partial\nu} = \dfrac{\partial\hat{v}_n}{\partial\nu} = \dfrac{\partial\hat{w}_n}{\partial\nu} = 0, & x \in \partial\Omega. \end{cases}$$

Similarly to the above, we can prove that there exist a subsequence of $\{(\hat{v}_n, \hat{w}_n)\}$, denoted by itself, and nonnegative functions $\hat{v}$ and $\hat{w}$, such that $(\hat{v}_n, \hat{w}_n) \to (\hat{v}, \hat{w})$ in $[C^{2+\alpha}(\bar{\Omega})]^2$ and $\|\hat{v}\|_\infty + \|\hat{w}\|_\infty = 1$. Moreover, if $\|v_n\|_\infty + \|w_n\|_\infty \geq \delta$ for some constant $\delta > 0$, then $(u, \hat{v}, \hat{w}, w)$ satisfies $\min_{\bar{\Omega}} \hat{v} = \min_{\bar{\Omega}} \hat{w} = 0$ and

$$\begin{cases} -d_1\Delta u = u(a - u - \varepsilon v - w), & x \in \Omega, \\ -\Delta\left(d_2\hat{v} + \dfrac{d_4\hat{v}}{\sigma + w^2}\right) = ku\hat{w} - \hat{v}, & x \in \Omega, \\ -d_3\Delta\hat{w} = b\hat{v} - m\hat{w}, & x \in \Omega, \\ \dfrac{\partial u}{\partial\nu} = \dfrac{\partial\hat{v}}{\partial\nu} = \dfrac{\partial\hat{w}}{\partial\nu} = 0, & x \in \partial\Omega. \end{cases} \tag{5.10}$$

If $\lim_{n\to\infty}\big(\|v_n\|_\infty + \|w_n\|_\infty\big) = 0$, then $v = w = 0$ and $(u, \hat{v}, \hat{w})$ satisfies

$$\begin{cases} -d_1\Delta u = u(a - u), & x \in \Omega, \\ -(d_2 + d_4/\sigma)\Delta\hat{v} = ku\hat{w} - \hat{v}, & x \in \Omega, \\ -d_3\Delta\hat{w} = b\hat{v} - m\hat{w}, & x \in \Omega, \\ \dfrac{\partial u}{\partial\nu} = \dfrac{\partial\hat{v}}{\partial\nu} = \dfrac{\partial\hat{w}}{\partial\nu} = 0, & x \in \partial\Omega. \end{cases} \tag{5.11}$$

In the case where (5.10) holds, writing the equation of $\hat{w}$ as

$$-d_3\Delta\hat{w} + m\hat{w} = b\hat{v} \geq 0 \ \text{ in } \ \Omega, \qquad \partial\hat{w}/\partial\nu = 0 \ \text{ on } \ \partial\Omega,$$

using $\min_{\bar{\Omega}} \hat{w} = 0$, and applying the strong maximum principle and the Hopf boundary lemma, we derive that $\hat{w} = 0$, and in turn $\hat{v} = 0$. This is a contradiction to $\|\hat{v}\|_\infty + \|\hat{w}\|_\infty = 1$.

When (5.11) holds, as $u > 0$, from the equation of u we conclude that $u = a$. Thus we have

$$\begin{cases} -(d_2 + d_4/\sigma)\Delta\hat{v} = ka\hat{w} - \hat{v}, & x \in \Omega, \\ -d_3\Delta\hat{w} = b\hat{v} - m\hat{w}, & x \in \Omega, \\ \dfrac{\partial\hat{v}}{\partial\nu} = \dfrac{\partial\hat{w}}{\partial\nu} = 0, & x \in \partial\Omega. \end{cases} \tag{5.12}$$

Since the parameters a, b, k, m are positive, and $\hat{v}$ and $\hat{w}$ are nonnegative and satisfy $\|\hat{v}\|_\infty + \|\hat{w}\|_\infty = 1$, by the strong maximum principle and the Hopf boundary lemma, we find that $\hat{v}$ and $\hat{w}$ are positive functions. Let $x_i, y_i \in \bar{\Omega}$ be such that $\hat{v}(x_1) = \min_{\bar{\Omega}} \hat{v}$,

$\hat{v}(y_1) = \max_{\bar{\Omega}} \hat{v}$, $\hat{w}(x_2) = \min_{\bar{\Omega}} \hat{w}$, and $\hat{w}(y_2) = \max_{\bar{\Omega}} \hat{w}$. Applying the maximum principle to (5.13), we have $\hat{v}(x_1) \geq ka\hat{w}(x_1) \geq ka\hat{w}(x_2)$, $\hat{v}(y_1) \leq ka\hat{w}(y_1) \leq ka\hat{w}(y_2)$, $m\hat{w}(x_2) \geq b\hat{v}(x_2) \geq b\hat{v}(x_1)$, $m\hat{w}(y_2) \leq b\hat{v}(y_2) \leq b\hat{v}(y_1)$. Since $\hat{v}(x_1) > 0$, it follows that $m = abk$, which is a contradiction to the condition $m < abk$.

Step 2. We now consider the remaining cases.

Since (u_n, v_n, w_n) satisfies $\int_\Omega v_n \mathrm{d}x = k \int_\Omega u_n w_n \mathrm{d}x$ and $m \int_\Omega w_n \mathrm{d}x = b \int_\Omega v_n \mathrm{d}x$, we have that

$$\int_\Omega v \mathrm{d}x = k \int_\Omega uw \mathrm{d}x, \quad m \int_\Omega w \mathrm{d}x = b \int_\Omega v \mathrm{d}x. \tag{5.13}$$

(i) If $d_1 = \infty$, then $u = u^* =$ constant. If $u^* = 0$, from (5.13), we have in turn that $v = w = 0$. This contradicts Lemma 1. So, $u^* > 0$.

(ia) If $d_2, d_3 < \infty$, then since either $\min_{\bar{\Omega}} v = 0$ or $\min_{\bar{\Omega}} w = 0$, similar to the arguments of Step 1, we have that $\min_{\bar{\Omega}} v = \min_{\bar{\Omega}} w = 0$. Note that the functions $\hat{v}_n, \hat{w}_n$ defined by (5.9) satisfy

$$\begin{cases} -\Delta\left(d_{2n}\hat{v}_n + \dfrac{d_{4n}\hat{v}_n}{\sigma + w_n^2}\right) = ku_n\hat{w}_n - \hat{v}_n, & x \in \Omega, \\ -d_{3n}\Delta\hat{w}_n = b\hat{v}_n - m\hat{w}_n, & x \in \Omega, \\ \dfrac{\partial \hat{v}_n}{\partial \nu} = \dfrac{\partial \hat{w}_n}{\partial \nu} = 0, & x \in \partial\Omega. \end{cases} \tag{5.14}$$

Similar to the above, we may assume that $(\hat{v}_n, \hat{w}_n) \to (\hat{v}, \hat{w})$ in $[C^{2+\alpha}(\bar{\Omega})]^2$ for some nonnegative functions $\hat{v}$ and $\hat{w}$, and $(\hat{v}, \hat{w})$ satisfies $\|\hat{v}\|_\infty + \|\hat{w}\|_\infty = 1$.

If $\|v_n\|_\infty + \|w_n\|_\infty \geq \delta$ for some constant $\delta > 0$, then $(\hat{v}, \hat{w})$ satisfies $\min_{\bar{\Omega}} \hat{v} = \min_{\bar{\Omega}} \hat{w} = 0$ and

$$\begin{cases} -\Delta\left(d_2\hat{v} + \dfrac{d_4\hat{v}}{\sigma + w^2}\right) = ku^*\hat{w} - \hat{v}, & x \in \Omega, \\ -d_3\Delta\hat{w} = b\hat{v} - m\hat{w}, & x \in \Omega, \\ \dfrac{\partial \hat{v}}{\partial \nu} = \dfrac{\partial \hat{w}}{\partial \nu} = 0, & x \in \partial\Omega. \end{cases}$$

Similar to the discussion of the problem (5.10) we can get a contradiction. If $\lim_{n\to\infty} \big(\|v_n\|_\infty + \|w_n\|_\infty\big) = 0$, then $v = w = 0$ and $(\hat{v}, \hat{w})$ satisfies

$$\begin{cases} -(d_2 + d_4/\sigma)\Delta\hat{v} = ku^*\hat{w} - \hat{v}, & x \in \Omega, \\ -d_3\Delta\hat{w} = b\hat{v} - m\hat{w}, & x \in \Omega, \\ \dfrac{\partial \hat{v}}{\partial \nu} = \dfrac{\partial \hat{w}}{\partial \nu} = 0, & x \in \partial\Omega. \end{cases} \tag{5.15}$$

As $(u, v, w) = (u^*, 0, 0)$, by Lemma 1, we have $u^* = a$. Thus, (5.15) is exactly (5.12). Similar to the arguments of the last part of Step 1, we arrive at $m = abk$—a contradiction.

Similarly, we can derive contradictions for all the other cases. □

6. Nonexistence of nonconstant positive solutions of (5.1) without cross diffusion. In this section we shall prove that, when $d_4 = 0$, the problem (5.1) has no nonconstant positive solution if d_1 is large. When $d_4 = 0$, (5.1) becomes

$$\begin{cases} -d_1\Delta u = au - u^2 - \varepsilon uv - uw, & x \in \Omega, \\ -d_2\Delta v = kuw - v, & x \in \Omega, \\ -d_3\Delta w = bv - mw, & x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x \in \partial\Omega. \end{cases} \tag{6.1}$$

The main result of this section is the following theorem.

THEOREM 4. *Let the parameters $d_2, d_3, a, b, k, m,$ and ε be fixed positive constants, and $m < abk$. Then there exists a positive constant $\hat{d}_1$ such that, when $d_1 \geq \hat{d}_1$,* (6.1) *has no nonconstant positive solutions.*

To prove this theorem, we will make use of the following lemma, which can be proved using results and methods of section 5. We shall omit the details.

LEMMA 2. *Let (u, v, u) be the positive solution of* (6.1). *Then we have*

$$\lim_{d_1 \to \infty} (u, v, w) = (\tilde{u}, \tilde{v}, \tilde{w}) \quad \textit{in } [C^2(\bar{\Omega})]^3, \tag{6.2}$$

where $(\tilde{u}, \tilde{v}, \tilde{w})$ is the positive constant solution of (6.1) *given by* (1.4).

Proof of Theorem 4. Define $W_\nu^{2,2}(\Omega) = \{u \in W^{2,2}(\Omega) : \frac{\partial u}{\partial \nu}|_{\partial\Omega} = 0\}$ and $W_{\nu,0}^{2,2}(\Omega) = W_\nu^{2,2}(\Omega) \cap L_0^2(\Omega)$, where $L_0^2(\Omega) = \{u \in L^2(\Omega) : \int_\Omega u \mathrm{d}x = 0\}$. Denote $\rho = d_1^{-1}$ and decompose $u = h + z$ with $h \in \mathbb{R}^1$ and $z \in W_{\nu,0}^{2,2}$. Let

$$F(\rho, h, z, v, w) = \begin{pmatrix} \displaystyle\int_\Omega (h+z)(a - h - z - \varepsilon v - w)\mathrm{d}x \\ \Delta z + \rho(h+z)(a - h - z - \varepsilon v - w) \\ d_2\Delta v - v + k(h+z)w \\ d_3\Delta w - mw + bv \end{pmatrix}.$$

Then $F : \mathbb{R}^2 \times W_{\nu,0}^{2,2}(\Omega) \times [W_\nu^{2,2}(\Omega)]^2 \to \mathbb{R}^1 \times L_0^2(\Omega) \times [L^2(\Omega)]^2$, and, for any $\rho > 0$, (u, v, w) solves (6.1) if and only if $F(\rho, h, z, v, w) = 0$. It is obvious that, for any ρ, we have $F(\rho, \tilde{u}, 0, \tilde{v}, \tilde{w}) = 0$.

Let Ψ be the Fréchet derivative of F at $(0, \tilde{u}, 0, \tilde{v}, \tilde{w})$ with respect to (h, z, v, w). A direct computation yields

$$\Psi(h, z, v, w) = \begin{pmatrix} \displaystyle -\tilde{u}\int_\Omega (h + z + \varepsilon v + w)\mathrm{d}x \\ \Delta z \\ d_2\Delta v - v + k\tilde{w}(h+z) + k\tilde{u}w \\ d_3\Delta w - mw + bv \end{pmatrix}.$$

We prove that Ψ is injective and surjective: It suffices to show that for any given $(h_1, z_1, v_1, w_1) \in \mathbb{R}^1 \times L_0^2(\Omega) \times [L^2(\Omega)]^2$, the equation $\Psi(h, z, v, w) = (h_1, z_1, v_1, w_1)$ has a unique solution $(h, z, v, w) \in \mathbb{R}^1 \times W_{\nu,0}^{2,2}(\Omega) \times [W_\nu^{2,2}(\Omega)]^2$, or equivalently, the following system has a unique solution:

$$\int_\Omega (h+z+\varepsilon v+w)\mathrm{d}x = -\frac{h_1}{\tilde{u}}, \tag{6.3}$$

$$\Delta z = z_1 \text{ in } \Omega, \quad \frac{\partial z}{\partial \nu}\Big|_{\partial\Omega} = 0, \quad \int_\Omega z\mathrm{d}x = 0, \tag{6.4}$$

$$\begin{cases} d_2\Delta v - v + k\tilde{w}(h+z) + (m/b)w = v_1, & x\in\Omega, \\ d_3\Delta w - mw + bv = w_1, & x\in\Omega, \\ \dfrac{\partial v}{\partial\nu} = \dfrac{\partial w}{\partial\nu} = 0, & x\in\partial\Omega. \end{cases} \tag{6.5}$$

Since $z_1 \in L_0^2(\Omega)$, (6.4) has a unique solution z. From (6.5) we have that

$$\begin{cases} \Delta(bd_2v + d_3w) + bk\tilde{w}(h+z) = bv_1 + w_1, & x\in\Omega, \\ \dfrac{\partial(bd_2v+d_3w)}{\partial\nu} = 0, & x\in\partial\Omega. \end{cases} \tag{6.6}$$

Since $\int_\Omega z\mathrm{d}x = 0$, (6.6) has a solution if and only if h satisfies

$$bk\tilde{w}h|\Omega| = \int_\Omega (bv_1 + w_1)\mathrm{d}x. \tag{6.7}$$

With such an h, which is obviously uniquely determined, (6.6) has a solution of the form $bd_2v + d_3w = g(x) + \lambda$, where λ is a constant (which will be uniquely determined later) and $g(x)$ is uniquely determined and satisfies $\int_\Omega g(x)\mathrm{d}x = 0$. The equation of w in (6.5) now becomes

$$d_3\Delta w - \left(m + \frac{d_3}{d_2}\right)w + \frac{g+\lambda}{d_2} = w_1 \text{ in } \Omega, \qquad \frac{\partial w}{\partial\nu} = 0 \text{ on } \partial\Omega.$$

For a given constant λ, this problem has a unique solution $w = w_\lambda(x)$ that satisfies

$$\int_\Omega w_1 dx + \left(m + \frac{d_3}{d_2}\right)\int_\Omega w\mathrm{d}x = \frac{1}{d_2}\int_\Omega g(x)\mathrm{d}x + \frac{\lambda}{d_2}|\Omega| = \frac{\lambda}{d_2}|\Omega|. \tag{6.8}$$

As a result, we have

$$\begin{aligned} \int_\Omega (\varepsilon v + w)\mathrm{d}x &= \int_\Omega \left(\varepsilon\frac{g(x)+\lambda-d_3w}{bd_2} + w\right)\mathrm{d}x \\ &= \frac{\lambda\varepsilon}{bd_2}|\Omega| + \left(1 - \frac{\varepsilon d_3}{bd_2}\right)\int_\Omega w\mathrm{d}x \\ &= \lambda\frac{(b+m\varepsilon)|\Omega|}{b(md_2+d_3)} - \frac{d_2}{md_2+d_3}\left(1-\frac{\varepsilon d_3}{bd_2}\right)\int_\Omega w_1 dx. \end{aligned}$$

Substituting this identity into (6.3) and making use of (6.7), we find that λ is uniquely determined by

$$-\frac{h_1}{\tilde{u}} - \frac{1}{bk\tilde{w}}\int_\Omega (bv_1+w_1)\mathrm{d}x = \lambda\frac{(b+m\varepsilon)|\Omega|}{b(md_2+d_3)} - \frac{d_2}{md_2+d_3}\left(1-\frac{\varepsilon d_3}{bd_2}\right)\int_\Omega w_1 dx,$$

and thus w and hence v are uniquely determined. In conclusion, for any given $(h_1, z_1, v_1, w_1) \in \mathbb{R}^1 \times L_0^2(\Omega) \times [L^2(\Omega)]^2$, the equation $\Psi(h,z,v,w) = (h_1,z_1,v_1,w_1)$

has a unique solution. This proves that Ψ is a one-to-one and surjective map between two Banach spaces. Therefore, Ψ^{-1} exists and is a bounded linear operator.

To complete the proof of Theorem 4, we note that, by the implicit function theorem, there is a constant $\delta > 0$ such that, for all $0 < \rho < \delta$, in a small neighborhood of $(\tilde{u}, 0, \tilde{v}, \tilde{w})$, the equation $F(\rho, h, z, v, w) = 0$ has a unique solution, which must be $(\tilde{u}, 0, \tilde{v}, \tilde{w})$. Correspondingly, when d_1 is large, in a small neighborhood of $(\tilde{u}, \tilde{v}, \tilde{w})$, the problem (6.1) has only the constant solution $(\tilde{u}, \tilde{v}, \tilde{w})$. This fact, combined with Lemma 2, concludes the proof. $\square$

7. Existence of stationary patterns for the PDE system with cross diffusion (1.6). In this section we shall discuss the existence of nonconstant positive solutions to (5.1). These solutions are not close to a constant solution and are obtained for large cross diffusion coefficient d_4, or for large diffusion coefficient d_1 (and $d_4 > 0$), with the other parameters $d_2, d_3, a, b, k, m, \varepsilon$, and σ suitably fixed. Those stationary patterns which are close to a positive constant solution are discussed in the next section, through bifurcation analysis. Our results here are as follows.

THEOREM 5. *Let the parameters $d_1, d_2, d_3, a, b, k, m, \varepsilon$, and σ be fixed such that $m < abk$ and $\sigma < \tilde{w}^2$. Let $\tilde{\mu}$ be given by the limit* (4.4). *If $\tilde{\mu} \in (\mu_n, \mu_{n+1})$ for some $n \geq 2$ and the sum $\sum_{i=2}^{n} \dim E(\mu_i)$ is odd, then there exists a positive constant d_4^* such that, for $d_4 \geq d_4^*$,* (5.1) *has at least one nonconstant positive solution.*

THEOREM 6. *Let the parameters $d_2, d_3, d_4, a, b, k, m, \varepsilon$, and σ be fixed so that $m < abk$ and* (4.6) *holds. Let $\bar{\mu}$ be given by the limit* (4.7). *If $\bar{\mu} \in (\mu_n, \mu_{n+1})$ for some $n \geq 2$ and the sum $\sum_{i=2}^{n} \dim E(\mu_i)$ is odd, then there exists a positive constant d_1^* such that, for $d_1 \geq d_1^*$,* (5.1) *has at least one nonconstant positive solution.*

Remark 3. The allowable values of $\tilde{\mu}$ and $\bar{\mu}$ in Theorems 5 and 6 may cover a wide range as the parameters vary. For example, $\tilde{\mu}$ is large when d_1 and d_3 are small while all the other parameters in Theorem 5 are fixed; $\bar{\mu}$ is large when d_3 is small and d_4 is large while all the other parameters in Theorem 6 are fixed and the required conditions are satisfied. Another way to see that the conditions $\tilde{\mu} \in (\mu_n, \mu_{n+1})$ and $\bar{\mu} \in (\mu_n, \mu_{n+1})$ are easily satisfied is to fix all the parameters but vary the underlying domain Ω. For example, if we replace Ω by $s\Omega = \{sx : x \in \Omega\}$, then μ_i is changed to $s^{-2}\mu_i$, and therefore the above conditions for $\tilde{\mu}$ and $\bar{\mu}$ are satisfied, respectively, for all s in a certain bounded interval.

Remark 4. Theorems 4 and 6 imply that when d_1 is large, stationary patterns arise only when cross diffusion is present, that is, $d_4 > 0$.

As the proofs of Theorems 5 and 6 are similar, we will prove only Theorem 6.

Proof of Theorem 6. By Proposition 3 and our assumption on $\bar{\mu}$, there exists a positive constant d_1^* such that, when $d_1 \geq d_1^*$, (4.8) holds and

$$\bar{\mu}_1(d_1) < 0 = \mu_1 < \bar{\mu}_2(d_1) < \mu_2, \quad \bar{\mu}_3(d_1) \in (\mu_n, \mu_{n+1}). \tag{7.1}$$

We shall prove that for any $d_1 \geq d_1^*$, (5.1) has at least one nonconstant positive solution. The proof, which is by contradiction, is based on the homotopy invariance of the topological degree.

Suppose on the contrary that the assertion is not true for some $d_1 = \bar{d}_1 \geq d_1^*$. In what follows we fix $d_1 = \bar{d}_1$.

For $t \in [0, 1]$, define $\Phi(t; \mathbf{u}) = \big([td_1 + (1-t)\hat{d}_1]u,\ d_2 v + t d_4 v/(\sigma + w^2),\ d_3 w\big)^T$, and consider the problem

$$\begin{cases} -\Delta\Phi(t; \mathbf{u}) = \mathbf{G}(\mathbf{u}), & x \in \Omega, \\ \dfrac{\partial \mathbf{u}}{\partial \nu} = 0, & x \in \partial\Omega, \end{cases} \tag{7.2}$$

where the positive constant $\hat{d}_1$ is determined by Theorem 4. Then $\mathbf{u}$ is a positive nonconstant solution of (5.1) if and only if it is such a solution of (7.2) for $t = 1$. It is obvious that $\tilde{\mathbf{u}}$ is the unique constant positive solution of (7.2) for any $0 \le t \le 1$. As we observed in section 4, for any $0 \le t \le 1$, $\mathbf{u}$ is a positive solution of (7.2) if and only if

$$\mathbf{F}(t;\,\mathbf{u}) \overset{\Delta}{=} \mathbf{u} - (\mathbf{I} - \Delta)^{-1}\Big\{\Phi_{\mathbf{u}}^{-1}(t;\,\mathbf{u})[\mathbf{G}(\mathbf{u}) + \nabla \mathbf{u}\,\Phi_{\mathbf{uu}}(t;\,\mathbf{u})\nabla \mathbf{u}] + \mathbf{u}\Big\} = 0 \quad \text{in} \quad \mathbf{Y}^+.$$

It is obvious that $\mathbf{F}(1;\,\mathbf{u}) = \mathbf{F}(\mathbf{u})$. Theorem 4 shows that $\tilde{\mathbf{u}}$ is the only solution of $\mathbf{F}(0;\,\mathbf{u}) = 0$ in $\mathbf{Y}^+$. By a direct computation, $D_{\mathbf{u}}\mathbf{F}(t;\,\tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\Phi_{\mathbf{u}}^{-1}(t;\,\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\}$. In particular, $D_{\mathbf{u}}\mathbf{F}(0;\,\tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\}$ and $D_{\mathbf{u}}\mathbf{F}(1;\,\tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\} = D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$, where $\mathcal{D} = \operatorname{diag}(\hat{d}_1, d_2, d_3)$. From (4.2) and (4.3) we see that

$$H(\mu) = \det\{\Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\}\mathcal{C}(\sigma, d_1, d_4; \mu). \tag{7.3}$$

In view of (4.8) and (7.1), it follows from (7.3) that

$$\begin{cases} H(\mu_1) = H(0) > 0, \\ H(\mu_i) < 0, \quad 2 \le i \le n, \\ H(\mu_i) > 0, \quad i \ge n+1. \end{cases}$$

Therefore, zero is not an eigenvalue of the matrix $\mu_i\mathbf{I} - \Phi_{\mathbf{u}}^{-1}(\tilde{\mathbf{u}})\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$ for all $i \ge 1$. Applying Proposition 1, we have that

$$\gamma = \sum_{i \ge 1, H(\mu_i) < 0} \dim E(\mu_i) = \sum_{i=2}^{n} \dim E(\mu_i), \quad \text{which is odd,}$$

and

$$\operatorname{index}(\mathbf{F}(1;\,\cdot),\,\tilde{\mathbf{u}}) = (-1)^\gamma = -1. \tag{7.4}$$

By Theorem 3 and its proof (where $B_3 > 0$), we can easily show that

$$\operatorname{index}(\mathbf{F}(0;\,\cdot),\,\tilde{\mathbf{u}}) = (-1)^0 = 1. \tag{7.5}$$

Now, by Propositions 6 and 7, there exists a positive constant C such that, for all $0 \le t \le 1$, the positive solutions of (7.2) satisfy $1/C < u,\, v,\, w < C$. Therefore, $\mathbf{F}(t;\,\mathbf{u}) \ne 0$ on $\partial B(C)$ for all $0 \le t \le 1$. By the homotopy invariance of the topological degree,

$$\deg(\mathbf{F}(1;\,\cdot),\,0,\,B(C)) = \deg(\mathbf{F}(0;\,\cdot),\,0,\,B(C)). \tag{7.6}$$

On the other hand, by our supposition, both equations $\mathbf{F}(1;\,\mathbf{u}) = 0$ and $\mathbf{F}(0;\,\mathbf{u}) = 0$ have only the positive solution $\tilde{\mathbf{u}}$ in $B(C)$, and hence, by (7.4) and (7.5), $\deg(\mathbf{F}(0;\,\cdot),\,0,\,B(C)) = \operatorname{index}(\mathbf{F}(0;\,\cdot),\,\tilde{\mathbf{u}}) = 1$ and $\deg(\mathbf{F}(1;\,\cdot),\,0,\,B(C)) = \operatorname{index}(\mathbf{F}(1;\,\cdot),\,\tilde{\mathbf{u}}) = -1$. This contradicts (7.6), and the proof is complete. □

8. Bifurcation. In this section, we discuss the bifurcation of nonconstant positive solutions of (5.1) with respect to the cross diffusion coefficient d_4 and the diffusion coefficient d_1.

In the consideration of bifurcation with respect to d_4, we recall that, for a constant solution $\mathbf{u}^*$, $(\tilde{d}_4;\, \mathbf{u}^*) \in (0,\, \infty) \times \mathbf{X}$ is a *bifurcation point* of (5.1) if, for any $\delta \in (0,\, \tilde{d}_4)$, there exists $d_4 \in [\tilde{d}_4 - \delta,\, \tilde{d}_4 + \delta]$ such that (5.1) has a nonconstant positive solution close to $\mathbf{u}^*$. Otherwise, we say that $(\tilde{d}_4;\, \mathbf{u}^*)$ is a *regular point*. Bifurcation and regular points with respect to d_1 are defined analogously.

We shall consider the bifurcation of (5.1) at the equilibrium points $(\tilde{d}_4;\, \tilde{\mathbf{u}})$, $\tilde{d}_4 > 0$, and $(\tilde{d}_1;\, \tilde{\mathbf{u}})$, $\tilde{d}_1 > 0$, respectively, while all other parameters are fixed. Let $\mathcal{S}_p = \{\mu_1,\, \mu_2,\, \mu_3,\, \ldots\}$ and $\Sigma = \{\mu > 0 \mid H(\mu) = 0\}$, where $H(\mu)$ is as defined in (4.2). To emphasize the dependence of $H(\mu)$ and Σ on d_4 or d_1, we write $H(d_4;\, \mu)$ or $H(d_1;\, \mu)$, and $\Sigma_{d_4}(d_4)$ or $\Sigma_{d_1}(d_1)$, respectively. We note that for each $d_4 > 0$ and $d_1 > 0$, Σ may have 0 or 2 elements.

The results of this section are contained in the following two theorems. Their proofs are based on the topological degree arguments used earlier in this paper. We shall omit them but refer the reader to similar treatments in [34].

THEOREM 7 (bifurcation with respect to d_4).

(1) *If* $\mathcal{S}_p \cap \Sigma_{d_4}(\tilde{d}_4) = \emptyset$, *then* $(\tilde{d}_4;\, \tilde{\mathbf{u}})$ *is a regular point of* (5.1).

(2) *Suppose* $\mathcal{S}_p \cap \Sigma_{d_4}(\tilde{d}_4) \neq \emptyset$ *and the positive roots of* $H(\tilde{d}_4;\, \mu) = 0$ *are all simple. If the number of elements in* $\mathcal{S}_p \cap \Sigma_{d_4}(\tilde{d}_4)$ *is odd, then* $(\tilde{d}_4;\, \tilde{\mathbf{u}})$ *is a bifurcation point of* (5.1). *In this case, there exists an interval* $(\alpha, \beta) \subset \mathbb{R}^+$, *where*

(i) $\tilde{d}_4 = \alpha < \beta < \infty$ *and* $\mathcal{S}_p \cap \Sigma_{d_4}(\beta) \neq \emptyset$, *or*

(ii) $0 < \alpha < \beta = \tilde{d}_4$ *and* $\mathcal{S}_p \cap \Sigma_{d_4}(\alpha) \neq \emptyset$, *or*

(iii) $(\alpha, \beta) = (\tilde{d}_4, \infty)$,

such that for every $d_4 \in (\alpha, \beta)$, (5.1) *admits a nonconstant positive solution.*

THEOREM 8 (bifurcation with respect to d_1).

(1) *If* $\mathcal{S}_p \cap \Sigma_{d_1}(\tilde{d}_1) = \emptyset$, *then* $(\tilde{d}_1;\, \tilde{\mathbf{u}})$ *is a regular point of* (5.1).

(2) *Suppose* $\mathcal{S}_p \cap \Sigma_{d_1}(\tilde{d}_1) \neq \emptyset$ *and the positive roots of* $H(\tilde{d}_1;\, \mu) = 0$ *are all simple. If the number of elements in* $\mathcal{S}_p \cap \Sigma_{d_1}(\tilde{d}_1)$ *is odd, then* $(\tilde{d}_1;\, \tilde{\mathbf{u}})$ *is a bifurcation point of* (5.1). *In this case, there exists an interval* $(c,\, d) \subset \mathbb{R}^+$, *where*

(i) $\tilde{d}_1 = c < d < \infty$ *and* $\mathcal{S}_p \cap \Sigma_{d_1}(d) \neq \emptyset$, *or*

(ii) $0 < c < d = \tilde{d}_1$ *and* $\mathcal{S}_p \cap \Sigma_{d_1}(c) \neq \emptyset$, *or*

(iii) $(c,\, d) = (\tilde{d}_1,\, \infty)$, *or*

(iv) $(c,\, d) = (0,\, \tilde{d}_1)$,

such that for every $d_1 \in (c,\, d)$, (5.1) *admits a nonconstant positive solution.*

9. Discussion. In this paper, we have introduced a more realistic mathematical model for a diffusive (spatially dependent) prey-predator system where the predator has a stage structure comprising immature and mature members. In this model, we have explicitly incorporated the interaction between the immature predator and the prey (through increased intake by the mature predators), and the interaction between the immature and the mature predator. In the latter interaction, we model the tendency of the immature predator to stay close to the mature predator by a cross diffusion. As a result, our model is a strongly coupled reaction-diffusion system, which is mathematically more complex than systems used to model stage-structured prey-predator behavior hitherto.

What is noteworthy about this model is that, as the cross diffusion term arises naturally as a reflection of the most salient feature of the immature-mature interaction (namely, that the immature tends to stay close to the mature within a species), it is precisely this cross diffusion that gives rise to stationary patterns for the model. Indeed, we have shown that stationary patterns do not arise for the ODE (spatially independent) model, nor the PDE model without cross diffusion.

We further remark that, besides capturing a salient biological behavior, this particular cross diffusion term is also significant from the mathematical point of view. Indeed, we are able to show that, of all the possible cross diffusions in (a simplified version of) this model (arising from various types of interactions between the different species and subspecies), only this cross diffusion term causes Turing instability. We therefore have the fortuitous situation where *the biological and mathematical interests converge on this cross diffusion term.*

To be more precise, the following represents the most general form of cross diffusions in the prey-predator model (1.5):

(9.1)
$$\begin{cases} u_t - \operatorname{div}\{K_{11}(\mathbf{u})\nabla u + K_{12}(\mathbf{u})\nabla v + K_{13}(\mathbf{u})\nabla w\} = G_1(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ v_t - \operatorname{div}\{K_{22}(\mathbf{u})\nabla v + K_{21}(\mathbf{u})\nabla u + K_{23}(\mathbf{u})\nabla w\} = G_2(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ w_t - \operatorname{div}\{K_{33}(\mathbf{u})\nabla w + K_{31}(\mathbf{u})\nabla u + K_{32}(\mathbf{u})\nabla v\} = G_3(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x \in \partial\Omega, \quad t > 0, \\ w(x,0) \geq 0, \ v(x,0) \geq 0, \ w(x,0) \geq 0, & x \in \Omega, \end{cases}$$

where $\mathbf{u}$ and $\mathbf{G}$ are as in (1.6), and biological considerations require that $K_{ij}(\mathbf{u})$ satisfy [28], [30, Ch.10]

$$\begin{cases} K_{11}(\mathbf{u}),\, K_{22}(\mathbf{u}),\, K_{33}(\mathbf{u}) > 0, \ \ K_{12}(\mathbf{u}),\, K_{13}(\mathbf{u}) \geq 0, \\ K_{21}(\mathbf{u}), \ \ K_{23}(\mathbf{u}), \ \ K_{31}(\mathbf{u}), \ \ K_{32}(\mathbf{u}) \leq 0. \end{cases} \tag{9.2}$$

A simpler version of (9.1) is the following:

$$\begin{cases} u_t - \Delta(d_1 u + u d_{12}(v) + u d_{13}(w)) = G_1(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ v_t - \Delta(d_2 v + v d_{21}(u) + v d_{23}(w)) = G_2(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ w_t - \Delta(d_3 w + w d_{31}(u) + w d_{32}(v)) = G_3(\mathbf{u}), & x \in \Omega, \quad t > 0, \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0, & x \in \partial\Omega, \quad t > 0, \\ w(x,0) \geq 0, \ v(x,0) \geq 0, \ w(x,0) \geq 0, & x \in \Omega, \end{cases}$$

where d_{ij} are nonnegative C^1 functions satisfying, according to (9.2), $d'_{12}(v)$, $d'_{13}(w) \geq 0$, $d'_{21}(u)$, $d'_{23}(w)$, $d'_{31}(u)$, $d'_{32}(v) \leq 0$.

Our analysis along the lines of this paper for each d_{ij} revealed that $d_{23}(w)$ has the most significant effect on the stability of $\tilde{\mathbf{u}}$. Indeed, except for $d_{23}(w)$, each of the other d_{ij} alone does not seem to cause instability of $\tilde{\mathbf{u}}$.

Finally, we also remark that, while our choice of the cross diffusion term is mainly mathematically motivated, as one of the simplest functions with the required property $d'_{23}(w) \leq 0$ and also giving rise to Turing instability and stationary patterns, it also reflects, as demonstrated earlier, the natural biological behavior of the immature and mature predators. As mentioned in section 1, a slightly more general form of cross diffusion can be adopted for essentially the same mathematical treatment.

Acknowledgments. The authors are grateful to the editor and the referees for helpful comments that greatly improved this paper.

REFERENCES

[1] H. Amann, *Dynamic theory of quasilinear parabolic equations* II: *Reaction-diffusion systems*, Differential Integral Equations, 3 (1990), pp. 13–75.

[2] S. M. Baer, B. W. Kooi, Yu. A. Kuznetsov, and H. R. Thieme, *Multiparametric bifurcation analysis of a basic two-stage population model*, SIAM J. Appl. Math., 66 (2006), pp. 1339–1365.

[3] V. Capasso and O. Diekmann, eds., *Mathematics Inspired by Biology*, Lecture Notes in Math. 1714, Springer-Verlag, Berlin, and CIME, Florence, 1999.

[4] V. Castets, E. Dulos, J. Boissonade, and P. DeKepper, *Experimental evidence of a sustained Turing-type equilibrium chemical pattern*, Phys. Rev. Lett., 64 (1990), pp. 2953–2956.

[5] J. Chattopadhyay and P. K. Tapaswi, *Effect of cross-diffusion on pattern formation: A nonlinear analysis*, Acta Appl. Math., 48 (1997), pp. 1–12.

[6] W. Y. Chen and R. Peng, *Stationary patterns created by cross-diffusion for the competitor-mutualist model*, J. Math. Anal. Appl., 291 (2004), pp. 550–564.

[7] X. F. Chen, Y. W. Qi, and M. X. Wang, *Steady states of a strongly coupled prey-predator model*, Discrete Contin. Dyn. Syst., suppl. (2005), pp. 173–180.

[8] X. F. Chen, Y. W. Qi, and M. X. Wang, *A strongly coupled predator-prey system with nonmonotonic functional response*, Nonlinear Anal., 67 (2007), pp. 1966–1979.

[9] K. L. Cooke, R. H. Elderkin, and W. Huang, *Predator-prey interactions with delays due to juvenile maturation*, SIAM J. Appl. Math., 66 (2006), pp. 1050–1079.

[10] M. C. Cross and P. S. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.

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

[12] Y. Du, *Order Structure and Topological Methods in Nonlinear Partial Differential Equations, Vol.* 1*: Maximum Principles and Applications*, World Scientific, Singapore, 2006.

[13] Y. Du and Y. Lou, *Qualitative behaviour of positive solutions of a predator-prey model: Effects of saturation*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 321–349.

[14] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equation of Second Order*, Springer-Verlag, Berlin, New York, 2001.

[16] S. A. Gourley and Y. Kuang, *A stage structured predator-prey model and its dependence on maturation delay and death rate*, J. Math. Biol., 49 (2004), pp. 188–200.

[17] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1993.

[18] D. Iron, M. J. Ward, and J. C. Wei, *The stability of spike solutions to the one-dimensional Gierer–Meinhardt model*, Phys. D, 150 (2001), pp. 25–62.

[19] Y. Kan-on and M. Mimura, *Singular perturbation approach to a 3-component reaction-diffusion system arising in population dynamics*, SIAM J. Math. Anal., 29 (1998), pp. 1519–1536.

[20] G. M. Lieberman, *Bounds for the steady-state Sel'kov model for arbitrary p in any number of dimensions*, SIAM J. Math. Anal., 36 (2005), pp. 1400–1406.

[21] C. S. Lin, W. M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis systems*, J. Differential Equations, 72 (1988), pp. 1–27.

[22] S.-Q. Liu and E. Beretta, *A stage-structured predator-prey model of Beddington–Deangelis type*, SIAM J. Appl. Math., 66 (2006), pp. 1101–1129.

[23] S. Q. Liu, L. S. Chen, and R. Agarwal, *Recent progress on stage-structured population dynamics*, Math. Comput. Modelling, 36 (2002), pp. 1319–1360.

[24] Y. Lou, S. Martinez, and W. M. Ni, *On* 3×3 *Lotka-Volterra competition systems with cross-diffusion*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 175–190.

[25] Y. Lou and W. M. Ni, *Diffusion, self-diffusion and cross-diffusion*, J. Differential Equations, 131 (1996), pp. 79–131.

[26] Y. Lou and W. M. Ni, *Diffusion vs cross-diffusion: An elliptic approach*, J. Differential Equations, 154 (1999), pp. 157–190.

[27] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1993.

[28] W. M. Ni, *Diffusion, cross-diffusion and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.

[29] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, AMS, Providence, RI, 2001.

[30] A. Okubo, *Diffusion and Ecological Problems: Mathematical Models*, Springer-Verlag, Berlin, New York, 1980.

[31] C. Ou and J. Wu, *Spatial spread of rabies revisited: Influence of age-dependent diffusion on nonlinear dynamics*, SIAM J. Appl. Math., 67 (2006), pp. 138–163.

[32] Q. Ouyang, R. Li, G. Li, and H. L. Swinney, *Dependence of Turing pattern wavelength on diffusion rate*, Notices J. Chem. Phys., 102 (1995), pp. 2551–2555.

[33] P. Y. H. Pang and M. X. Wang, *Qualitative analysis of a ratio-dependent predator-prey system with diffusion*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 919–942.

[34] P. Y. H. Pang and M. X. Wang, *Nonconstant positive steady states of a predator-prey system with nonmonotonic functional response and diffusion*, Proc. London Math. Soc., 88 (2004), pp. 135–157.

[35] P. Y. H. Pang and M. X. Wang, *Strategy and stationary pattern in a three-species predator-prey model*, J. Differential Equations, 200 (2004), pp. 245–274.

[36] R. Peng and M. X. Wang, *Positive steady-state solutions of the Noyes–Field model for Belousov–Zhabotinskii reaction*, Nonlinear Anal., 56 (2004), pp. 451–464.

[37] A. Turing, *The chemical basis of morphogenesis*, Philos. Trans. R. Soc. London B, 237 (1952), pp. 37–72.

[38] M. X. Wang, *Nonconstant positive steady states of the Sel'kov model*, J. Differential Equations, 190 (2003), pp. 600–620.

[39] M. X. Wang, *Stationary patterns of strongly coupled predator-prey models*, J. Math. Anal. Appl., 292 (2004), pp. 484–505.

[40] M. X. Wang, *Stationary patterns for a prey-predator model with prey-dependent and ratio-dependent functional responses and diffusion*, Phys. D, 196 (2004), pp. 172–192.

[41] W. D. Wang and L. S. Chen, *A predator-prey system with stage-structure for predator*, Comput. Math. Appl., 33 (1997), pp. 83–91.

[42] W. D. Wang, G. Mulone, F. Salemi, and V. Salone, *Permanence and stability of a stage-structured predator-prey model*, J. Math. Anal. Appl., 262 (2001), pp. 499–528.

[43] X.-F. Wang, *Qualitative behavior of solutions of chemotactic diffusion systems: Effects of motility and chemotaxis and dynamics*, SIAM J. Math. Anal., 31 (2000), pp. 535–560.

[44] Y. N. Xiao and L. S. Chen, *Global stability of a predator-prey system with stage structure for the predator*, Acta Math. Sin. (Engl. Ser.), 20 (2004), pp. 63–70.

[45] R. Xu, M. A. J. Chaplain, and F. A. Davidson, *Persistence and global stability of a ratio-dependent predator-prey model with stage structure*, Appl. Math. Comput., 158 (2004), pp. 729–744.

[46] E. Yanagida, *Mini-maximizer for reaction-diffusion systems with skew-gradient structure*, J. Differential Equations, 179 (2002), pp. 311–335.

SIAM J. Appl. Math.
Vol. 69, No. 2, pp. 621–639

COEXISTENCE OF LIMIT CYCLES AND HOMOCLINIC LOOPS IN A SIRS MODEL WITH A NONLINEAR INCIDENCE RATE[*]

YILEI TANG[†], DEQING HUANG[‡], SHIGUI RUAN[§], AND WEINIAN ZHANG[‡]

Abstract. Recently, Ruan and Wang [*J. Differential Equations*, 188 (2003), pp. 135–163] studied the global dynamics of a SIRS epidemic model with vital dynamics and a nonlinear saturated incidence rate. Under certain conditions they showed that the model undergoes a Bogdanov–Takens bifurcation; i.e., it exhibits saddle-node, Hopf, and homoclinic bifurcations. They also considered the existence of none, one, or two limit cycles. In this paper, we investigate the coexistence of a limit cycle and a homoclinic loop in this model. One of the difficulties is to determine the multiplicity of the weak focus. We first prove that the maximal multiplicity of the weak focus is 2. Then feasible conditions are given for the uniqueness of limit cycles. The coexistence of a limit cycle and a homoclinic loop is obtained by reducing the model to a universal unfolding for a cusp of codimension 3 and studying degenerate Hopf bifurcations and degenerate Bogdanov–Takens bifurcations of limit cycles and homoclinic loops of order 2.

Key words. degenerate Bogdanov–Takens bifurcation, degenerate Hopf bifurcation, limit cycle, homoclinic loop, revised sign list

AMS subject classifications. 34C23, 92D30

DOI. 10.1137/070700966

1. Introduction.

Periodic oscillations are common phenomena observed in the incidence of many infectious diseases such as chickenpox, influenza, measles, mumps, rubella, etc. (see Hethcote [10, 11], Hethcote and Levin [12], Hethcote, Stech, and van den Driessche [13]). It is very important to understand such epidemic patterns in order to introduce public health interventions and control the spread of diseases. Recent studies have demonstrated that the incidence rate plays a crucial role in producing periodic oscillations in epidemic models (Alexander and Moghadas [1, 2], Derrick and van den Driessche [6], Hethcote and van den Driessche [14], Liu et al. [17, 18], Lizana and Rivero [19], Moghadas [21], Moghadas and Alexander [22], Ruan and Wang [25], Wang [26]).

In most epidemic models (see Anderson and May [3]), the *incidence rate* (the number of new cases per unit time) takes the mass-action form with bilinear interactions, namely, $\kappa S(t)I(t)$, where $S(t)$ and $I(t)$ are the numbers of susceptible and infectious individuals at time t, respectively, and the constant κ is the probability of transmission per contact. Epidemic models with such bilinear incidence rates usually have at most one endemic equilibrium and do not exhibit periodicity; the disease will be eradicated if the basic reproduction number is less than one and will persist otherwise (Anderson and May [3], Hethcote [11]). There are many reasons for using nonlinear incidence rates, and various forms of nonlinear incidence rates have

[*] Received by the editors August 23, 2007; accepted for publication (in revised form) August 5, 2008; published electronically December 3, 2008.
http://www.siam.org/journals/siap/69-2/70096.html

[†] Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China and Department of Mathematics, Sichuan University, Chengdu, Sichuan 610064, China (mathtyl@163.com).

[‡] Department of Mathematics, Sichuan University, Chengdu, Sichuan 610064, China (toglyhdq@sina.com, matzwn@126.com). The work of the fourth author was supported by NSFC (China) grants 10571127 and 10825104 and SRFDP 20050610003.

[§] Department of Mathematics, University of Miami, Coral Gables, FL 33124-4250 (ruan@math.miami.edu). This author's research was supported by NSF grant DMS-0715772.

been proposed recently. For example, in order to incorporate the effect of behavioral changes, Liu, Levin, and Iwasa [18] used a nonlinear incidence rate of the form

$$g(I)S = \frac{\kappa I^{\ell} S}{1 + \alpha I^{h}}, \tag{1.1}$$

where κI^{ℓ} measures the infection force of the disease, $1/(1+\alpha I^h)$ describes the inhibition effect from the behavioral change of the susceptible individuals when the number of infectious individuals increases, ℓ, h, and κ are all positive constants, and α is a nonnegative constant. See also Alexander and Moghadas [1, 2], Derrick and van den Driessche [6], Hethcote and van den Driessche [14], Moghadas [21], etc. Notice that the bilinear interaction is a special case of (1.1) with $\alpha = 0$ and $\ell = 1$.

The nonlinear function $g(I)$ given by (1.1) includes three types. (a) Unbounded incidence function: $\ell > h$. The case when $\ell = h + 1$ was considered by Hethcote and van den Driessche [14]. The function is unbounded as the bilinear incidence rate (see Figure 1(a)). (b) Saturated incidence function: $\ell = h$. The case when $\ell = h = 1$, i.e., $g(I) = \kappa I/(1+\alpha I)$, was proposed by Capasso and Serio [5] to describe a "crowding effect" or "protection measures" in modeling the cholera epidemics in Bari in 1973. A similar type of sigmoidal function was also used to represent dose-response relationships observed in parasite infection experiments (Regoes, Ebert, and Bonhoeffer [23]). The function tends to a saturation level as the number of infectious individuals I becomes large (see Figure 1(b)). (c) Nonmonotone incidence function: $\ell < h$. Such functions can be used to interpret the "psychological effects" (Capasso and Serio [5]): for a very large number of infectious individuals the infection force may decrease as the number of infectious individuals increases (see Figure 1(c)), because in the presence of a large number of infectious individuals the population may tend to reduce the number of contacts per unit time, as seen with the spread of SARS (see Wang [26], Xiao and Ruan [28]).

From the graphs in Figure 1, one would expect that the dynamics of epidemic models with unbounded incidence rates are similar to those with bilinear incidence rates. In fact, Hethcote and van den Driessche [14] found that in a SEIRS model with $\ell = h + 1$, the classical threshold results hold; namely, the disease dies out below the threshold, and the disease level approaches the endemic equilibrium above the threshold. For a SIRS model with the nonmonotone incidence function $g(I) = kI/(1 + \alpha I^2)$, Xiao and Ruan [28] demonstrated that either the number of infectious individuals tends to zero as time evolves or the disease persists. We conjecture that the dynamics of SIRS models with nonmonotone incidence rates are similar to those observed by Xiao and Ruan [28].

On the other hand, the dynamics of epidemic models with saturated incidence rates (when $\ell = h$) have been shown to be very rich and complex. For a SEIRS model with $\ell = h$, Hethcote and van den Driessche [14] observed that the threshold concept becomes more complicated since the asymptotic behavior can depend on both the threshold and the initial values. The model can have none, one, or two endemic equilibria, and the disease can die out above the threshold for some initial values. Periodic solutions appear through Hopf bifurcation. The results are analogous to those obtained by Liu, Hethcote, and Levin [17] for SIRS models with $\ell = h$. The case $\ell = h = 1$ has been discussed briefly by Capasso and Serio [5] and recently in some detail by Gomes et al. [8], who obtained the existence of backward bifurcations, oscillations, and Bogdanov–Takens points in SIR and SIS models. These indicate that the case when $\ell = h \geq 2$ can be very complicated and deserves further investigation.

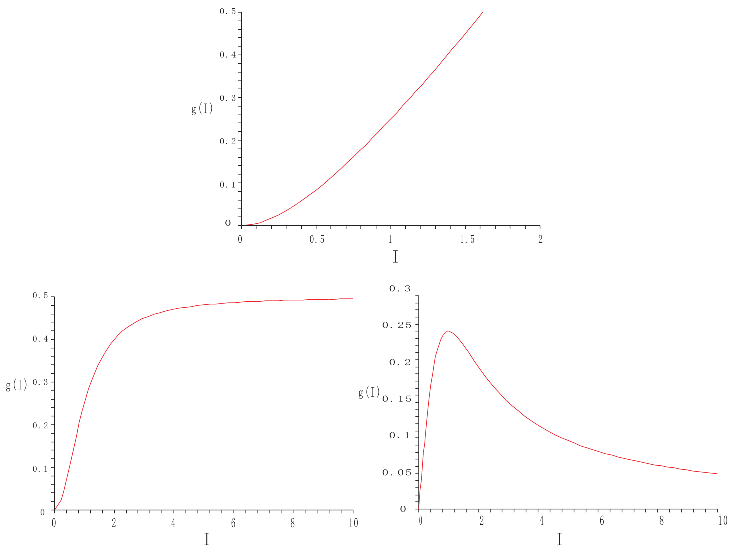


FIG. 1. *Graphs of the incidence function* $g(I) = \frac{\kappa I^\ell}{1+\alpha I^h}$. (a) *Top: Unbounded incidence when* $\ell > h$ *(*$\ell = 2, h = 1$*).* (b) *Bottom-left: Saturated incidence when* $\ell = h$ *(*$\ell = h = 2$*).* (c) *Bottom-right: Nonmonotone incidence when* $\ell < h$ *(*$\ell = 1, h = 2$*). Here,* $k = 0.5, \alpha = 1$.

In order to better understand the generic bifurcations in SIRS models with saturated incidence rates and also motivated by the work of Liu et al. [17, 18] and Hethcote and van den Driessche [14], Ruan and Wang [25] studied the global dynamics of a SIRS model with the nonlinear incidence function $g(I) = \kappa I^2/(1 + \alpha I^2)$, i.e., $\ell = h = 2$:

$$
(1.2) \qquad \begin{cases} \dfrac{dS}{dt} &= b - \delta S - \dfrac{\kappa I^2 S}{1 + \alpha I^2} + \nu R, \\ \dfrac{dI}{dt} &= \dfrac{\kappa I^2 S}{1 + \alpha I^2} - (\delta + \gamma) I, \\ \dfrac{dR}{dt} &= \gamma I - (\delta + \nu) R, \end{cases}
$$

where $b > 0$ is the recruitment rate of the population, $\delta > 0$ is the death rate of the population, $\gamma > 0$ is the recovery rate of infectious individuals, and $\nu > 0$ is the rate of removed individuals who lose immunity and return to the susceptible class. Summing up the three equations in (1.2), we obtain an equation $dN/dt = b - \delta N$ with $N(t) = S(t) + I(t) + R(t)$. Obviously, all solutions of this equation tend to its equilibrium $N(t) \equiv N_0 = b/\delta$ as $t \to +\infty$. Thus, all important dynamical behaviors of system (1.2) occur on the plane $S + I + R = N_0$, and the restricted system on the plane becomes

$$
(1.3) \qquad \begin{cases} \dfrac{dI}{dt} = \dfrac{\kappa I^2}{1 + \alpha I^2}(N_0 - I - R) - (\delta + \gamma) I, \\ \dfrac{dR}{dt} = \gamma I - (\delta + \nu) R. \end{cases}
$$

Under certain conditions Ruan and Wang [25] showed that the simplified model (1.3) undergoes a Bogdanov–Takens bifurcation; i.e., it exhibits saddle-node, Hopf, and homoclinic bifurcations. They also established the existence of none, one, or two limit cycles by applying the Bendixson–Dulac criterion [32], the Poincaré–Bendixson theorem [9], and a classic method for uniqueness of limit cycles in the Liénard equation [32], respectively. The coexistence (Theorem 2.9 in [25]) of two limit cycles is obtained by assuming that the successor function [32] (denoted by d in [25]) can switch its signs.

In Ruan and Wang [25] the uniqueness of limit cycles was obtained under the assumption that a polynomial $h(x)$ of degree 6 is nonpositive for all x in a definite interval, which is actually not easy to check. Moreover, only the first order Liapunov value of the weak focus (I_2, R_2) in Theorem 2.6 of [25] was calculated. To have a better understanding of the dynamics of the system, we need to calculate higher order Liapunov values of the weak focus, which is difficult in general. In fact, the weak focus (I_2, R_2) in Theorem 2.6 of [25] (E_+ in this paper) may have multiplicity 2, two limit cycles may arise from a degenerate Hopf bifurcation, and a limit cycle and a homoclinic loop may coexist via the degenerate Bogdanov–Takens bifurcation.

The study of model (1.3) is interesting and significant since it exhibits different and complicated dynamics such as periodic solutions, homoclinic orbits, multiple endemic equilibria, etc. The global dynamics is still not well understood. In this paper we further study the dynamical behavior of system (1.3). By rescaling the variables

$$x = \left(\sqrt{\frac{\kappa}{\delta+\nu}}\right) I, \quad y = \left(\sqrt{\frac{\kappa}{\delta+\nu}}\right) R, \quad d\tau = (\delta+\nu)dt/(1+pI^2)$$

and parameters

$$p = \frac{\alpha(\delta+\nu)}{\kappa}, \quad A = N_0\sqrt{\frac{\kappa}{\delta+\nu}}, \quad m = \frac{\delta+\gamma}{\delta+\nu}, \quad q = \frac{\gamma}{\delta+\nu},$$

system (1.3) is transformed into an equivalent system

$$(1.4) \qquad \begin{cases} \dfrac{dI}{dt} &= -I[(mp+1)I^2 + (R-A)I + m] &=: \mathcal{I}(I,R), \\ \dfrac{dR}{dt} &= (1+pI^2)(qI-R) &=: \mathcal{R}(I,R), \end{cases}$$

where we still use I, R, t to present x, y, τ for simplicity and $I, R \geq 0$, $A, m, p, q > 0$. We first calculate the second order Liapunov value at the weak focus and prove that the maximal multiplicity of the weak focus is 2 by technically dealing with some complicated multivariable polynomials, which implies that at most two limit cycles can arise near the weak focus. Then, by reducing the determination of the sign for polynomials of higher degrees to revised sign lists [31], we give some clean conditions on the parameters for the uniqueness of limit cycles. Finally, we reduce system (1.4) to a form of universal unfolding for a cusp of codimension 3 so as to give the bifurcation surfaces and display all limit cycles and homoclinic loops of order up to 2, from which the coexistence of limit cycles and homoclinic loops is established.

The paper is organized as follows. Some preliminary results on the existence and properties of equilibria are reviewed in section 2. Section 3 is devoted to the study of degenerate Hopf bifurcation. The uniqueness of limit cycles is considered in section 4. In section 5, we study the degenerate Bogdanov–Takens bifurcation of the model. A brief discussion on the models, motivations, methods, and results is given in section 6.

2. Preliminaries.

We first recall some known results on the existence of equilibria. As shown in Ruan and Wang [25], system (1.4) has at most three equilibria $O = (0,0)$, $E_- = (I_-, R_-)$, and $E_+ = (I_+, R_+)$ in the first quadrant, where

$$I_\pm = \frac{A \pm (A^2 - 4m(mp+q+1))^{1/2}}{2(mp+q+1)}, \quad R_\pm = qI_\pm.$$

It is easy to see that O is the disease-free equilibrium of system (1.4) and is a stable node. Moreover, there are no positive equilibria if $A^2 < 4m(mp+q+1)$ and two positive ones E_- and E_+ if $A^2 > 4m(mp+q+1)$. They coincide at $E_0 = (I_0, R_0) = (A/2(mp+q+1), qA/2(mp+q+1))$ if $A^2 = 4m(mp+q+1)$. It is indicated in [25] that E_- is a saddle and E_+ is a node, a focus, or a center. Moreover, the following results are given in Theorem 2.1 in [25].

Lemma 2.1. *The equilibrium E_+ is stable if one of the following inequalities holds:*

$$A^2 > A_c^2, \;\; m \le 1, \;\; q < \frac{2mp+1}{m-1},$$

where

$$A_c^2 =: \frac{(mq+2m-1-q+2m^2p)^2}{(m-1)(mp+p+1)}.$$

E_+ is unstable if

$$A^2 < A_c^2, \;\; m > 1, \;\; \text{and} \;\; q > \frac{2mp+1}{m-1}.$$

When the parameters lie in the region

$$\Omega = \{(A,m,p,q) |\, m > 1,\, q > (2mp+1)/(m-1),\, A^2 = A_c^2\}, \tag{2.1}$$

the linearization of system (1.4) at E_+ has a pair of purely imaginary eigenvalues. Let

$$\mu = (1+2m-q(m-1)) + (4+2m+4q-6mq+6m^2+2m^2q)p + 4m(m^2+2)p^2. \tag{2.2}$$

The following results on Hopf bifurcation are given in Theorem 2.6 in [25].

Lemma 2.2. *Suppose that conditions in Ω hold. If $\mu < 0$, then there is a stable periodic orbit in* (1.4) *as A^2 decreases from A_c^2. If $\mu > 0$, there is an unstable periodic orbit in* (1.4) *as A^2 increases from A_c^2. If $\mu = 0$, a Hopf bifurcation with codimension* 2 *may occur.*

Obviously, in [25] a question remains open: Is E_+ possibly a center when $\mu = 0$? A negative answer will be given in section 3. Regarding μ as a quadratic polynomial of p, we can easily see that the case $\mu = 0$ happens if and only if the discriminant of (2.2) is ≥ 0.

As shown previously, when $A = A_0 =: 2\sqrt{m(mp+q+1)}$, the equilibrium E_0 appears in the interior of the first quadrant and is degenerate because the Jacobian matrix of the linearized system of (1.4) at E_0 has determinant 0.

Lemma 2.3. *When $A = A_0$, E_0 is either a saddle-node if $p \ne ((m-1)q-1)/(2m)$ or a cusp otherwise.*

Proof. For $p = ((m-1)q-1)/(2m)$ it was proved in [25] that system (1.4) has a cusp at E_0. Consider the case that $p \ne ((m-1)q-1)/(2m)$. With the change of variables $(I,R) \mapsto (x,y)$ defined by

$$x = -\frac{(1+q+2mp)(I-I_0)}{1+q+mp} + \frac{m(R-R_0)}{1+q+mp}, \quad y = -\frac{mq(I-I_0)}{1+q+mp} + \frac{m(R-R_0)}{1+q+mp},$$

system (1.4) is rewritten as

$$
\text{(2.3)}\quad \left\{\begin{array}{rcl}
\dot{x} &=& \frac{-d_0\sqrt{1+mp+q}\sqrt{m}}{(b_0q-d_0)^2}x^2 + \frac{2(pqb_0^2+b_0d_0+b_0d_0mp-b_0pd_0+d_0^2)\sqrt{m}}{b_0\sqrt{1+mp+q}(b_0q-d_0)^2}xy \\
&& -\frac{(2pqb_0^2-b_0d_0q+b_0d_0-2b_0pd_0+b_0d_0mp+2d_0^2)\sqrt{m}}{b_0\sqrt{1+mp+q}(b_0q-d_0)^2}y^2 + O(|(x,y)|^3) \quad =: X_1(x,y), \\
\dot{y} &=& (d_0-b_0q)y - \frac{b_0q\sqrt{1+mp+q}\sqrt{m}}{(b_0q-d_0)^2}x^2 + \frac{2(b_0q+qmpb_0+qpb_0+qd_0-pd_0)\sqrt{m}}{\sqrt{1+mp+q}(b_0q-d_0)^2}xy \\
&& -\frac{(-b_0q^2+b_0q+qmpb_0+2qpb_0+2qd_0-2pd_0)\sqrt{m}}{\sqrt{1+mp+q}(b_0q-d_0)^2}y^2 + O(|(x,y)|^3) \qquad =: Y_1(x,y),
\end{array}\right.
$$

where $b_0 = -m/(1+q+mp)$, $d_0 = -(1+q+2mp)/(1+q+mp)$, and E_0 is translated to the origin. By the implicit function theorem, there is a unique function $y = \varsigma(x)$ such that $\varsigma(0) = 0$ and $Y_1(x, \varsigma(x)) = 0$. Actually, we can solve from $Y_1(x,y) = 0$ that

$$
\varsigma(x) = -\frac{qb_0\sqrt{1+q+mp}\sqrt{m}}{(b_0q-d_0)^3}x^2 + O(|x|^3).
$$

Substituting $y = \varsigma(x)$ into the first equation of (2.3), we get

$$
\text{(2.4)}\qquad \dot{x} = X_1(x, \varsigma(x)) = -\frac{d_0\sqrt{1+q+mp}\sqrt{m}}{(b_0q-d_0)^2}x^2 + O(|x|^3).
$$

Theorem 7.1 in Chapter 2 of [32] implies that the origin is a saddle-node of system (2.3). Thus, E_0 is a saddle-node of system (1.4). □

3. Degenerate Hopf bifurcation. This section is a complement to the Hopf bifurcation analysis in Ruan and Wang [25]. In Lemma 2.2 the sign of μ is the same as the sign of the first Liapunov value of (1.4) if E_+ is a weak focus, but [25] does not determine the sign of the higher order Liapunov values and whether E_+ is a center. In this section, we overcome some technical difficulties in the computation of the higher order Liapunov values and prove that E_+ is a weak focus of multiplicity at most 2.

As in section 2, we consider those parameters in the region Ω, defined in (2.1), where the Jacobian matrix at E_+ has a pair of purely imaginary eigenvalues and the parameters A, m, p, and q satisfy $A^2 = A_c^2 > 4m(mp+q+1)$. In this case, the first coordinate of E_+ takes the form $I_+ = (2m^2p - q + mq - 1 + 2m)/(A(mp+p+1))$. A simple transformation $(I, R) \mapsto (x, y)$, which translates E_+ to the origin and diagonalizes the linear part, reduces system (1.4) to

$$
\text{(3.1)}\qquad \left\{\begin{array}{rcl}
\frac{dx}{dt} &=& -wy + ma_1x^2 + 2a_1wxy - ma_1^2x^3 - a_1^2wx^2y, \\
\frac{dy}{dt} &=& wx - a_1b_2x^2 - 2a_1xy + a_1^2b_1x^3 + a_1^2x^2y,
\end{array}\right.
$$

where

$$
\begin{aligned}
w &= \sqrt{k_1}, \quad k_1 = -\frac{(2mp+1)(2mp-mq+1+q)}{(mp+p+1)^2}, \quad a_1 = \frac{mp+p+1}{\sqrt{(m-1)(mp+p+1)}}, \\
b_1 &= \frac{2m^3p^2+pqm^2+3m^2p+2mp^2-2pqm+m+p+qp}{(mp+p+1)^2w}, \\
b_2 &= \frac{2m^3p^2-2m^2p^2+3m^2p+2qm^2p+4mp^2-4pqm-mp+m+2qp+2p}{(mp+p+1)^2w}.
\end{aligned}
$$

Obviously, $k_1 > 0$ because $m > 1$ and $q > (2mp+1)/(m-1)$. Using the polar coordinates $x_1 = r\cos\theta$, $y_1 = r\sin\theta$, we obtain from (3.1) that

$$\begin{aligned}\frac{dr}{d\theta} &= \tfrac{G_2(\theta)}{w}r^2 + (\tfrac{G_3(\theta)}{w} - \tfrac{G_2(\theta)H_1(\theta)}{w^2})r^3 + (-\tfrac{G_3(\theta)H_1(\theta)}{w^2} - \tfrac{G_2(\theta)H_2(\theta)}{w^2} + \tfrac{G_2(\theta)H_1^2(\theta)}{w^3})r^4 \\ &\quad +(-\tfrac{G_3(\theta)H_2(\theta)}{w^2} + \tfrac{2G_2(\theta)H_1(\theta)H_2(\theta)}{w^3} + \tfrac{G_3(\theta)H_1^2(\theta)}{w^3} - \tfrac{G_2(\theta)H_1^3(\theta)}{w^4})r^5 + \text{h.o.t.},\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}G_2(\theta) &= a_1(m+2)\sin^3\theta + a_1(2w-b_2)\sin^2\theta\cos\theta - 2a_1\sin\theta,\\ G_3(\theta) &= -a_1^2(m+1)\sin^4\theta + a_1^2(b_1-w)\sin^3\theta\cos\theta + a_1^2\sin^2\theta,\\ H_1(\theta) &= a_1(2w-b_2)\sin^3\theta - a_1(m+2)\sin^2\theta\cos\theta - 2a_1w\sin\theta,\\ H_2(\theta) &= a_1^2(b_1-w)\sin^4\theta + a_1^2(m+1)\sin^3\theta\cos\theta + a_1^2w\sin^2\theta,\end{aligned}$$

and

$$G_4(\theta) = G_5(\theta) = H_3(\theta) = H_4(\theta) = H_5(\theta) = 0.$$

Consider solutions of (3) in the formal series $r(\theta, r_0) = \sum_{j=1}^{+\infty} r_j(\theta)r_0^j$ together with the initial condition $r(0, r_0) = r_0$, where $|r_0|$ is sufficiently small. Obviously, $r_1(0) = 1$, $r_2(0) = r_3(0) = \cdots = 0$. Substituting the series into (3) and comparing the coefficients, we obtain a system of differential equations for $r_j(\theta), j = 1, 2, \ldots$, i.e.,

$$\frac{dr_1}{d\theta} = 0, \quad \frac{dr_2}{d\theta} = r_1^2\frac{G_2(\theta)}{w}, \quad \frac{dr_3}{d\theta} = r_1^3\left(\frac{G_3(\theta)}{w} - \frac{G_2(\theta)H_1(\theta)}{w^2}\right) + 2r_1r_2\frac{G_2(\theta)}{w}, \ldots. \tag{3.3}$$

Solving them together with the initial conditions, we get

$$\begin{aligned}r_1(\theta) \equiv 1, \quad r_2(\theta) &= \int_0^\theta \frac{G_2(\xi)}{w}d\xi,\\ r_3(\theta) &= \int_0^\theta \left\{\frac{G_3(\xi)}{w} - \frac{G_2(\xi)H_1(\xi)}{w^2} + \frac{2G_2(\xi)r_2(\xi)}{w}\right\}d\xi, \ldots.\end{aligned}$$

Using Maple V.7 software, we compute the Liapunov value as follows:

$$L_3(m, w, a_1, b_1, b_2) = \frac{1}{2\pi}r_3(2\pi) = \frac{a_1^2(m-1)(2b_2-w)}{8w^2}. \tag{3.4}$$

For parameters in Ω, the sign of L_3 is determined by $2b_2 - w$ and therefore is the same as the sign of μ, which demonstrates the corresponding results in Lemma 2.2.

If $w = 2b_2$ (i.e., $\mu = 0$), then $L_3(m, w, a_1, b_1, b_2) = 0$. In this case the Liapunov value of order 5 can be calculated as follows:

$$L_5(m, w, a_1, b_1, b_2) = \frac{1}{2\pi}r_5(2\pi) = \frac{a_1^4(m-1)\kappa(m, w, a_1, b_1, b_2)}{768w^4} > 0, \tag{3.5}$$

where, using (3.1) and the fact that $\mu = 0$, we have

$$\begin{aligned}\kappa(m, w, a_1, b_1, b_2) &= 121w^3 - 332w^2b_2 + 143w^2b_1 - 75wm^2 - 218wm - 190wb_2b_1\\ &\quad -8w + 155wb_2^2 + 150b_2m^2 + 436mb_2 + 16b_2 + 50b_2^3\\ &= \frac{(2p+1)(mp+p+1)^3(-4mp+8p+2)^{1/2}}{64(m-1)(2mp+1)^2(m^2p+mp+m)^{1/2}} > 0.\end{aligned}$$

By the theory of Hopf bifurcation [20, 32], we obtain the following results.

THEOREM 3.1. *Suppose that $A^2 > 4m(mp + q + 1)$ and conditions in Ω hold.*

(i) *If $\mu \neq 0$, then the equilibrium E_+ of system* (1.4) *is a weak focus of multiplicity* 1 *and at most one limit cycle arises from the Hopf bifurcation. Moreover, E_+ is stable and the limit cycle is also stable when $\mu < 0$, or E_+ is unstable and the limit cycle is also unstable when $\mu > 0$.*

(ii) *If $\mu = 0$, then the equilibrium E_+ is a weak focus of multiplicity* 2 *and at most two limit cycles arise from the Hopf bifurcation. Moreover, E_+ is unstable and the outer cycle is also unstable, but the inner cycle (if it appears) is stable.*

To carry out numerical simulations on two limit cycles, we choose parameters $m = 3$, $p = 0.1$, $q = 15$, and $A = 21.99$. We can verify that the conditions of result (ii) in Theorem 3.1 are satisfied. In this case the two limit cycles can be simulated by using MATLAB 6.5 software. In Figure 2, the trajectory started at the point $P_1 = (1, 5)$ spirals inward as the time $t \to -\infty$ and the trajectory started at $P_2 = (1.2, 12)$ spirals inward as $t \to \infty$. Hence, an unstable outer limit cycle exists and lies in the annular region bounded by these two trajectories. Similarly, the orbit started at $P_3 = (1.5, 15)$ spirals inward as $t \to \infty$ and the orbit started at $P_4 = (1.3, 17)$ spirals outward as $t \to \infty$. Therefore, there is a stable inner limit cycle lying in the annular region bounded by these two orbits. The equilibrium $E_+ = (1.195073225, 17.92609838)$ is unstable.

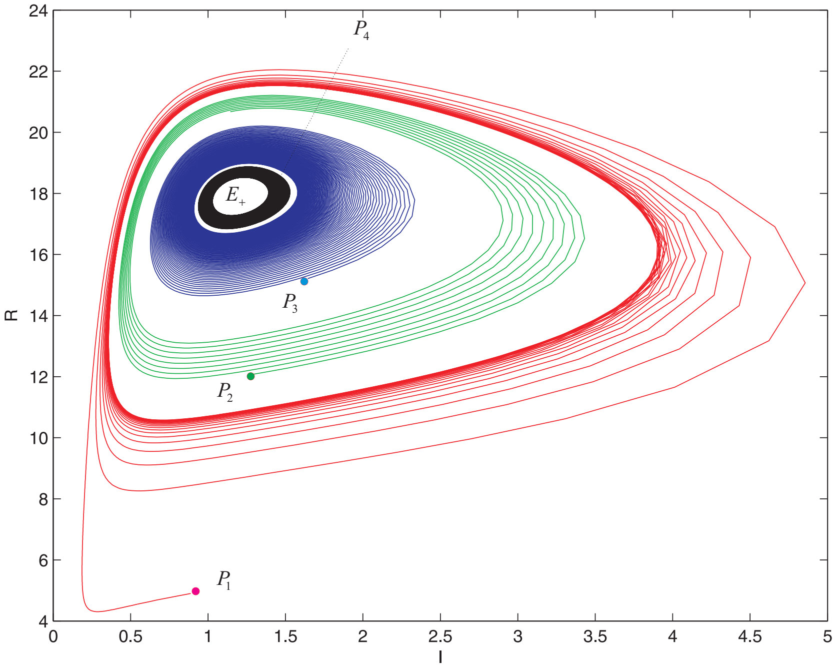


FIG. 2. *Two limit cycles bifurcate from the weak focus of multiplicity* 2.

4. Uniqueness of limit cycles. In this section we consider the uniqueness of limit cycles of system (1.4) and provide a relatively simpler proof compared to that of [25]. As shown in Theorem 2.2 of [25] and section 2, it suffices to discuss the case when $m > 1$ and $A^2 > 4m(mp + q + 1)$, in which system (1.4) possibly has closed orbits and E_+ lies in the first quadrant.

Our strategy is to reduce system (1.4) to the form of the Liénard system

$$\dot{x} = y - F(x), \qquad \dot{y} = -g(x) \tag{4.1}$$

and apply the known Theorem 1.1 in Kooij and Zegeling [15] and Theorem 2.1 in Xiao and Zhang [29]. Rearranging terms in the order of powers of R, we rewrite system (1.4) as

$$\dot{I} = g_0(I) - g_1(I)R, \qquad \dot{R} = qI - R, \tag{4.2}$$

where $g_0(I) = (A - I)I^2/(1 + pI^2) - mI$ and $g_1(I) = I^2/(1 + pI^2)$. We need only to consider $I > 0$ because $I = 0$ is an orbit. Thus $g_1(I) > 0$, and system (4.2) has the same phase portrait as the system

$$\dot{I} = R - \frac{g_0(I)}{g_1(I)}, \qquad \dot{R} = \frac{R}{g_1(I)} - \frac{qI}{g_1(I)}. \tag{4.3}$$

With the transformation

$$I = x, \qquad y = R - \int_{I_+}^{x} \frac{1}{g_1(x)} dx, \tag{4.4}$$

system (4.3) is reduced to the Liénard system (4.1) with

$$F(x) = \frac{g_0(x)}{g_1(x)} - \int_{I_+}^{x} \frac{1}{g_1(x)} dx, \qquad g(x) = \frac{qx}{g_1(x)} - \frac{g_0(x)}{g_1^2(x)}. \tag{4.5}$$

LEMMA 4.1. *Suppose that $m > 1$ and $A^2 > 4m(mp + q + 1)$. Then system* (4.2) *has at most one closed orbit in the interior of the first quadrant if either the function $F'(x)/g(x)$, where F' denotes the derivative of F, is neither decreasing nor a constant or $F'(I_+)\frac{d}{dx}(F'(x)/g(x)) < 0$ for $x \neq I_+$. Moreover, the closed orbit is hyperbolic if it exists.*

Proof. Note that limit cycles of system (4.2) (if any exist) lie in the stripe region between the vertical lines $\ell_0 : I = I_-$ and $\ell_1 : I = I_D$, where

$$I_D = \frac{A + (A^2 - 4m(mp + 1))^{1/2}}{2(mp + 1)}.$$

In fact, a vertical isocline of (4.2) intersects the I-axis at $D = (I_D, 0)$, i.e., $(mp + 1)I_D^2 - AI_D + m = 0$. Restricted to ℓ_1 in the interior of the first quadrant, the derivative $\dot{I}$ satisfies

$$\dot{I}|_{\ell_1} = I_D\{[-m + AI_D - (mp + 1)I_D^2] - RI_D\} = -RI_D^2 < 0.$$

This implies that limit cycles of system (4.2) (if any exist) lie on the left of ℓ_1 because $I_+ < I_D$ (i.e., the equilibrium E_+ lies on the left of ℓ_1). On the other hand, limit cycles lie on the right of ℓ_0; otherwise, a limit cycle intersects ℓ_0 because $I_+ > I_-$ (i.e., E_+ lies on the right of ℓ_0), implying that $\dot{I} = 0$ at a point on ℓ_0. This is a contradiction because on ℓ_0 the derivative

$$\dot{I} = I_-\{[-m + (A - R_-)I_- - (mp + 1)I_-^2] + R_-I_- - RI_-\} = (R_- - R)I_-^2 \neq 0$$

except at the saddle E_-. Let $S(I_-, I_D)$ denote the stripe region. Since transformations between (4.2) and (4.1) do not change x, it suffices to discuss (4.1) in $S(I_-, I_D)$, i.e., $I_- < x < I_D$.

Transformation (4.4) is one-to-one for $I > 0$ and $R > 0$, so it is equivalent to discuss the uniqueness of closed orbits for system (4.1) where $x > 0$. Corresponding to E_+, system (4.1) has an equilibrium (x_+, y_+) with the same coordinates, i.e., $x_+ = I_+$ and $y_+ = R_+$. From (4.5) we have $F'(x) = -\{(mp + p + 1)x^2 + 1 - m\}/x^2$ and $g(x) = \rho(x)(1 + px^2)/x^3$, where $\rho(x) = (mp + q + 1)x^2 - Ax + m$ has exactly two zeros at I_- and I_+. It follows that

$$g(x_+) = -\frac{g_0(I_+) - g_1(I_+)R_+}{g_1^2(I_+)} = 0, \quad (x - x_+)g(x) = \left(\frac{1 + px^2}{x^3}\right)(x - x_+)\rho(x) > 0$$

in $S(I_-, I_D)$, verifying partly either the condition in [29, Theorem 2.1] (as $F'(x)/g(x)$ is neither decreasing nor a constant) or the condition in [15, Theorem 1.1] (as $F'(I_+) \cdot (d/dx)\{F'(x)/g(x)\} < 0$ for $x \neq I_+$). The other conditions can be checked explicitly by the assumptions in our lemma. Thus the lemma is proved. □

In order to obtain conditions in terms of the original parameters for the uniqueness of limit cycles and complete the results of the uniqueness in [25], we use the notation

$$c_0 = m(m-1), \quad c_2 = 5m^2p + mp + 4m + mq - q - 1, \quad c_3 = 2A(1 + 2mp),$$
$$c_4 = (5m^2p^2 - mp^2) + (4mpq - 2pq) + (6mp - 2p) + q + 1, \quad c_6 = p(1+p+mp)(1+q+mp).$$

These constants are obviously all positive. Using the conditions in Lemma 4.1, we obtain the following theorem.

THEOREM 4.2. *Suppose that $m > 1$ and $A^2 > 4m(mp + q + 1)$. If either* (i) $c_3^2 - 4c_2c_4 > 0$ *and $\sigma_- < I_-$, $\sigma_+ > I_D$, where $\sigma_\pm = (c_3 \pm \sqrt{c_3^2 - 4c_2c_4}/(2c_4)$, or* (ii) $c_3^2 - 4c_2c_4 < 0$, *$F'(I_+) > 0$, and both c_6 and $m(m-1)$ are small enough, then system* (4.2) *has at most one limit cycle in the interior of the first quadrant. Moreover, under assumption* (i) *(resp.,* (ii)*) the limit cycle is unstable (resp., stable) if it exists.*

Proof. Calculate $(d/dx)\{F'(x)/g(x)\} = h(x)/x^6g^2(x)$ in $S(I_-, I_D)$, where

$$h(x) = c_0 - c_2x^2 + c_3x^3 - c_4x^4 + c_6x^6.$$

By Lemma 4.1, we need to determine the sign of $h(x)$. Our strategy is to discuss the quadratic function $\eta(x) = -c_2 + c_3x - c_4x^2$, for which we have $h(x) = m(m-1) + c_6x^6 + \eta(x)x^2$.

In case (i), the function η has exactly two real roots $\sigma_\pm$ and $[I_-, I_D] \subset (\sigma_-, \sigma_+)$. Since η has the leading coefficient $c_4 > 0$, we see that $\eta(x) > 0$ for $x \in (\sigma_-, \sigma_+)$, i.e., $\eta(x) > 0$ in $S(I_-, I_D)$. Since $m > 1$, as required in (2.1), we have $h(x) > 0$ or equivalently $(d/dx)\{F'(x)/g(x)\} > 0$ in $S(I_-, I_D)$. Then Lemma 4.1 implies the result in case (i).

In case (ii), $c_3^2 - 4c_2c_4 < 0$, so $\eta(x) \neq 0$. On the compact interval $[I_-, I_D]$ the function $|\eta|$ is bounded by a positive number. Therefore, $\mathrm{sgn} h(x) = \mathrm{sgn} \eta(x)$ for $x \in (I_-, I_D)$ as c_6 and $m(m-1)$ are both sufficiently small. Since η has the leading coefficient $c_4 > 0$, we ensure that $\eta(x) < 0$ as $x \in (I_-, I_D)$, implying that $h(x) < 0$ in $S(I_-, I_D)$. Thus, $F'(I_+)(d/dx)\{F'(x)/g(x)\} < 0$; i.e., the second condition in Lemma 4.1 holds in $S(I_-, I_D)$. From the notion introduced at the beginning of the proof of Lemma 4.1, it suffices to verify the condition in $S(I_-, I_D)$. Therefore, the conclusion in case (ii) is obtained. □

Consider $p = q = \epsilon^k$, $m = 1/\epsilon$, and $(2A(2mp+1))^2 = a_1c_2c_4$, for example, where $\epsilon > 0$ is sufficiently small, $k \geq 3$, and $4 < a_1 < 9/2$. These parameters satisfy assumption (i) in Theorem 4.2. Another choice of parameters in which $p = \epsilon^{2k}$, $m = 1+\epsilon$, and $q = 1/\epsilon^k$, where $\epsilon > 0$ is sufficiently small and $k \geq 3$, verifies assumption (ii) in Theorem 4.2.

For c_6 and $m(m-1)$ to be sufficiently small in case (ii) of Theorem 4.2, we have to restrict p and m near 0 and 1, respectively. Efforts are also made to extend the restriction by some known results on the zeros of high degree polynomials (Yang [31]). As shown in the above proof for Theorem 4.2 (ii), we can generally suppose that $m > 1$, $A^2 > 4m(mp+q+1)$ and that $h(I_-) < 0$, $h(I_D) < 0$ and claim that h has no real zeros in the interval (I_-, I_D). By Lemma 3.1 in [31], the number of real zeros of h in (I_-, I_D) is equal to the number of negative zeros of the function

$$\Psi(x) = (1-x)^6 h\left(\frac{I_D - I_- x}{1-x}\right) = \alpha_0 x^6 + \alpha_1 x^5 + \alpha_2 x^4 + \alpha_3 x^3 + \alpha_4 x^2 + \alpha_5 x + \alpha_6,$$

where

$$\begin{aligned}
\alpha_0 &= h(I_-) = c_0 - c_2 I_-^2 + c_3 I_-^3 - c_4 I_-^4 + c_6 I_-^6,\\
\alpha_1 &= -3c_3 I_D I_-^2 + 4c_4 I_D I_-^3 - 6c_6 I_D I_-^5 + 2c_2 I_D I_- + 2c_4 I_-^4 - 6c_0 + 4c_2 I_-^2 - 3c_3 I_-^3,\\
\alpha_2 &= 9c_3 I_D I_-^2 + 3c_3 I_D^2 I_- - 8c_4 I_D I_-^3 - 6c_4 I_D^2 I_-^2 + 15c_6 I_D^2 I_-^4 - 8c_2 I_D I_- - c_4 I_-^4\\
&\quad +15c_0 - c_2 I_D^2 - 6c_2 I_-^2 + 3c_3 I_-^3,\\
\alpha_3 &= -9c_3 I_D I_-^2 - 9c_3 I_D^2 I_- + 4c_4 I_D I_-^3 + 12c_4 I_D^2 I_-^2 + 4c_4 I_D^3 I_-\\
&\quad -20c_6 I_D^3 I_-^3 + 12c_2 I_D I_- - 20c_0 + 4c_2 I_D^2 + 4c_2 I_-^2 - c_3 I_-^3 - c_3 I_D^3,\\
\alpha_4 &= 3c_3 I_D I_-^2 + 9c_3 I_D^2 I_- - 6c_4 I_D^2 I_-^2 - 8c_4 I_D^3 I_- + 15c_6 I_D^4 I_-^2 - 8c_2 I_D I_- + 15c_0\\
&\quad -c_4 I_D^4 - 6c_2 I_D^2 - c_2 I_-^2 + 3c_3 I_D^3,\\
\alpha_5 &= -3c_3 I_D^3 - 3c_3 I_D^2 I_- - 6c_6 I_D^5 I_- + 4c_2 I_D^2 + 2c_2 I_D I_- + 2c_4 I_D^4 + 4c_4 I_D^3 I_- - 6c_0,\\
\alpha_6 &= h(I_D) = c_0 - c_2 I_D^2 + c_3 I_D^3 - c_4 I_D^4 + c_6 I_D^6.
\end{aligned}$$

Let Discr(Ψ) be the *discrimination matrix* of the polynomial Ψ, constructed in the appendix as in [31, Definition 2.1] and its following paragraph, and calculate its principal minors $\varpi_1, \varpi_2, \ldots, \varpi_{13}$ as in the appendix. Consider the sequence $SE = \{\varpi_1\varpi_2, \varpi_2\varpi_3, \ldots, \varpi_{12}\varpi_{13}\}$ and its sign list $\mathcal{S}(SE) = \{\mathrm{sgn}(\varpi_1\varpi_2), \ldots, \mathrm{sgn}(\varpi_{12}\varpi_{13})\}$, where $\mathrm{sgn}(x)$ denotes the sign of x. Now revise the signs according to the following rule (Definition 2.3 in [31]): (S1) If $\{\varrho_i, \varrho_{i+1}, \ldots, \varrho_{i+j}\}$ is a section of $\mathcal{S}(SE)$ such that $\varrho_i \neq 0, \varrho_{i+1} = \cdots = \varrho_{i+j-1} = 0, \varrho_{i+j} \neq 0$, then replace the section with the finite sequence $\{-\varrho_i, -\varrho_i, \varrho_i, \varrho_i, -\varrho_i, -\varrho_i, \varrho_i, \varrho_i, \ldots\}$ by truncating for the same number of terms; (S2) otherwise, do not change. Let $\mathcal{S}'(SE)$ denote the *revised sign list.* By Theorem 3.3 in [31], the number of distinct negative zeros of Ψ is equal to $\xi_1 - \xi$, where ξ is the number of sign changes in $\mathcal{S}'(SE)$ and $2\xi_1$ is the number of nonzero members in $\mathcal{S}'(SE)$. Thus we conclude that Ψ *has no negative zeros, i.e.,* $h(x) < 0$ *in* (I_-, I_D) *if* $\xi = \xi_1$.

The above conclusion shows that the condition on parameters for $h(x) < 0$ can be determined by the list $\mathcal{S}'(SE)$. It is easy to calculate that $\varpi_1\varpi_2 = 6\alpha_0^3 < 0$, i.e., $\mathrm{sgn}(\varpi_1\varpi_2) = -1$. So in total we have 3^{11}(=177147) cases to discuss because each of the remaining 11 elements in $\mathcal{S}'(SE)$ has three options: $-1, 0, 1$. We illustrate a general method for conditions on parameters with a further discussion on $\varpi_{12}\varpi_{13}$. In the case that $\varpi_{12}\varpi_{13} \neq 0$, $\mathcal{S}'(SE)$ contains 12 nonzero members, implying that $\xi_1 = 6$. So we need only to construct a revised sign list with $\xi = 6$. We easily find such a list $\{-1,1,-1,1,-1,1,-1,-1,-1,-1,-1,-1\}$, which gives a condition on parameters:

$$\begin{aligned}(\mathrm{C}_1):&\varpi_2\varpi_3\geq 0, \varpi_3\varpi_4<0, \varpi_4\varpi_5>0, \varpi_5\varpi_6<0, \varpi_6\varpi_7>0, \varpi_7\varpi_8<0,\\ &\varpi_8\varpi_9<0, \varpi_9\varpi_{10}<0, \varpi_{10}\varpi_{11}<0, \varpi_{11}\varpi_{12}<0, \varpi_{12}\varpi_{13}<0.\end{aligned}$$

In the case that $\varpi_{12}\varpi_{13}=0$, the number of nonzero members in $\mathcal{S}'(SE)$ is <12, i.e., $\xi_1\leq 5$. Note that the list $\{-1,1,-1,1,-1,1,1,1,1,1,0,0\}$ has $\xi=5$. Being a revised sign list, it gives a condition of parameters:

$$\begin{aligned}(\mathrm{C}_2):&\varpi_2\varpi_3\geq 0, \varpi_3\varpi_4<0, \varpi_4\varpi_5>0, \varpi_5\varpi_6<0, \varpi_6\varpi_7>0, \varpi_7\varpi_8>0,\\ &\varpi_8\varpi_9>0, \varpi_9\varpi_{10}>0, \varpi_{10}\varpi_{11}>0, \varpi_{11}\varpi_{12}=\varpi_{12}\varpi_{13}=0.\end{aligned}$$

Finally, Lemma 4.1 and the conclusion given in the last paragraph enable us to summarize that if $m>1$, $A^2>4m(mp+q+1)$, $h(I_-)<0$, $h(I_D)<0$, $F'(I_+)>0$, and if either (C_1) or (C_2) holds, then system (4.2) has at most one limit cycle in the interior of the first quadrant. Moreover, the limit cycle is stable if it exists. More conditions other than (C_1) and (C_2) can similarly be obtained for the uniqueness of limit cycles.

5. Degenerate Bogdanov–Takens bifurcation. By Lemma 2.3, when $A=A_0=:2\sqrt{m(mp+q+1)}$ and $p=p_0=:((m-1)q-1)/(2m)$, the equilibrium $E_0=:(A_0/(2(mp_0+q+1)), qA_0/(2(mp_0+q+1)))$ is a cusp, where the Bogdanov–Takens bifurcation may occur by a perturbation. By the standard theory of the Bogdanov–Takens bifurcation (of codimension 2), Ruan and Wang [25] assert only that the system has at most *one* limit cycle and the obtained homoclinic loop is of order 1 (see the definition in [16]).

In the following, we display the possible bifurcations of multiple limit cycles and homoclinic loops of order higher than 1. Note that as $A=A_0$ and $p=p_0$, we have $(m-1)q=2mp_0+1>0$, implying $m>1$. So we fix $m_0>1$ near 1 arbitrarily and consider three bifurcation parameters A,p,m near A_0,p_0,m_0, respectively. Let

$$A=A_0+\epsilon_1,\quad p=p_0+\epsilon_2,\quad m=m_0+\epsilon_3, \tag{5.1}$$

where $\epsilon_3>0$. Then, we discuss bifurcations of the equivalent system (1.4) for the parameters $\epsilon=(\epsilon_1,\epsilon_2,\epsilon_3)$ near $(0,0,0)$.

Lemma 5.1. *For A,p,m close to A_0,p_0,m_0, respectively,* (1.4) *is equivalent to the system*

$$\frac{dx}{dt}=y,\quad \frac{dy}{dt}=\mu_1+x^2+(\mu_2+\mu_3x+x^3+O(|x|^4))y+G(x,\mu)y^2, \tag{5.2}$$

where μ_i's are functions of $\epsilon_1,\epsilon_2,\epsilon_3$ such that $\frac{\partial(\mu_1,\mu_2,\mu_3)}{\partial(\epsilon_1,\epsilon_2,\epsilon_3)}|_{\epsilon=0}\neq 0$ and $G(x,\mu)$ is a C^∞ function.

Proof. With the substitution (5.1), equation (1.4) can be written as

$$\begin{cases}\dot{I}=\mathcal{I}(I+\frac{A_0}{2((m_0+\epsilon_3)(p_0+\epsilon_2)+q+1)}, R+\frac{qA_0}{2((m_0+\epsilon_3)(p_0+\epsilon_2)+q+1)}),\\ \dot{R}=\mathcal{R}(I+\frac{A_0}{2((m_0+\epsilon_3)(p_0+\epsilon_2)+q+1)}, R+\frac{qA_0}{2((m_0+\epsilon_3)(p_0+\epsilon_2)+q+1)}),\end{cases} \tag{5.3}$$

where $\mathcal{I}$ and $\mathcal{R}$ are defined in (1.4). When $\epsilon=0$, system (5.3) has a cusp at the origin $O_2=(0,0)$, as shown in [25]; i.e., the equilibrium E_0 is translated to O_2. Expanding (5.3) at O_2, rescaling time by $t=\tau(q+1+qm_0)/(2qm_0)$, and then applying a linear transformation T_1: $(I,R)\mapsto(\tilde{I},\tilde{R})$, defined by $\tilde{I}=I$ and $\tilde{R}=I-R/q$ to reduce the matrix of the linear part for $\epsilon=0$ to the Jordan canonical form, we can reduce (5.3) further to the form

$$\begin{bmatrix}\dot{\tilde{I}}\\ \dot{\tilde{R}}\end{bmatrix}=\begin{bmatrix}\vartheta_1(\epsilon)\\ \vartheta_2(\epsilon)\end{bmatrix}+\begin{bmatrix}\iota_{11}(\epsilon) & \iota_{12}(\epsilon)\\ \iota_{21}(\epsilon) & \iota_{22}(\epsilon)\end{bmatrix}\begin{bmatrix}\tilde{I}\\ \tilde{R}\end{bmatrix}+\begin{bmatrix}\omega_1(\tilde{I},\tilde{R},\epsilon)\\ \omega_2(\tilde{I},\tilde{R},\epsilon)\end{bmatrix}, \tag{5.4}$$

where all ϑ_j,ι_{ij}, and ω_j $(i,j=1,2)$ are calculated as in the appendix, which satisfies that $\vartheta_1(0)=\vartheta_2(0)=0$, $\iota_{12}(0)=1$, and $\iota_{11}(0)=\iota_{21}(0)=\iota_{22}(0)=0$. Another transformation T_2: $(\tilde{I},\tilde{R})\mapsto(X,Y)$, defined by

$$\tilde{I}=X+\frac{\sqrt{2m_0(q+1+qm_0)}}{2m_0}X^2,\quad \tilde{R}=Y+\frac{\sqrt{2m_0(q+1+qm_0)}(q+1+qm_0)}{4qm_0}X^2, \tag{5.5}$$

reduces system (5.4) to

$$\dot{X}=E_{11}(X,\epsilon)+E_{12}(X,\epsilon)Y,\quad \dot{Y}=E_{21}(X,\epsilon)+E_{22}(X,\epsilon)Y, \tag{5.6}$$

where

$$\begin{aligned}E_{1j}(X,\epsilon)&= a_{j0}(\epsilon)+a_{j1}(\epsilon)X+a_{j2}(\epsilon)X^2+a_{j3}(\epsilon)X^3+O(|X^4|),\quad j=1,2,\\ E_{2j}(X,\epsilon)&= b_{j0}(\epsilon)+b_{j1}(\epsilon)X+b_{j2}(\epsilon)X^2+b_{j3}(\epsilon)X^3+O(|X^4|),\quad j=1,2,\end{aligned}$$

and all $a_{ij}(\epsilon)$'s and $b_{ij}(\epsilon)$'s are given in the appendix. Applying the change of variables $\tilde{X}=X$, $\tilde{Y}=E_{11}(X,\epsilon)+E_{12}(X,\epsilon)Y$ in (5.6), we obtain a system in which the first equation is same as the first equation of (5.2), that is,

$$\dot{\tilde{X}}=\tilde{Y},\quad \dot{\tilde{Y}}=F_1(\tilde{X},\epsilon)+F_2(\tilde{X},\epsilon)\tilde{Y}+F_3(\tilde{X},\epsilon)\tilde{Y}^2, \tag{5.7}$$

where $F_i(\tilde{X},\epsilon)=\sum_{j=0}^{3}c_{ij}(\epsilon)\tilde{X}^j+O(|\tilde{X}|^4)$, $i=1,2$, and both $c_{ij}(\epsilon)$'s and $F_3(\tilde{X},\epsilon)$ are given in the appendix. Note that $c_{12}(0)=-\sqrt{2m_0(q+1+qm_0)}(q+1+qm_0)/4qm_0$ <0, which implies that $c_{12}(\epsilon)<0$ for small ϵ, and it is reasonable to apply the rescaling $\tilde{X}\mapsto-\tilde{X}$, $\tilde{Y}\mapsto-\sqrt{-c_{12}(\epsilon)}\,\tilde{Y}$, $\tau\mapsto\tau/\sqrt{-c_{12}(\epsilon)}$ to system (5.7) and obtain

$$\dot{\tilde{X}}=\tilde{Y},\quad \dot{\tilde{Y}}=\tilde{F_1}(\tilde{X},\epsilon)+\tilde{F_2}(\tilde{X},\epsilon)\tilde{Y}+\tilde{F_3}(\tilde{X},\epsilon)\tilde{Y}^2, \tag{5.8}$$

where $\tilde{F_1}(\tilde{X},\epsilon)=c_{10}(\epsilon)/c_{12}(\epsilon)-(c_{11}(\epsilon)/c_{12}(\epsilon))\tilde{X}+\tilde{X}^2+O(|\tilde{X}|^3)$, $\tilde{F_2}(\tilde{X},\epsilon)=\{c_{20}(\epsilon)-c_{21}(\epsilon)\tilde{X}+c_{22}(\epsilon)\tilde{X}^2-c_{23}(\epsilon)\tilde{X}^3\}/\sqrt{-c_{12}(\epsilon)}+O(|\tilde{X}|^4)$, and $\tilde{F_3}(\tilde{X},\epsilon)=-F_3(-\tilde{X},\epsilon)$. Thus, the coefficient of $\tilde{X}^2$ in $\tilde{F_1}$ in the second equation of (5.8) reduces to 1, the same as the corresponding one in (5.2).

In order to reduce system (5.8) to the induced form (5.2), we need to remove the term of $\tilde{X}$ in the second equation of (5.8). We achieve this by the affine transformation $u=\tilde{X}-c_{11}(\epsilon)/2c_{12}(\epsilon)$, $v=\tilde{Y}$ in $\tilde{X}$, and change system (5.8) into

$$\dot{u}=v,\quad \dot{v}=G_1(u,\epsilon)+G_2(u,\epsilon)v+G_3(u,\epsilon)v^2, \tag{5.9}$$

where $G_1(u,\epsilon)=d_{10}(\epsilon)+u^2+O(|u|^3,\epsilon^2)$, $G_2(u,\epsilon)=d_{20}(\epsilon)+d_{21}(\epsilon)u+d_{22}(\epsilon)u^2+d_{23}(\epsilon)u^3+O(|u|^4)$, $G_3(u,\epsilon)=-F_3(-u-c_{11}/(2c_{12}),\epsilon)$, and the coefficients $d_{ij}(\epsilon)$ are displayed in the appendix. Because

$$d_{23}(0) = \frac{2^{-\frac{3}{4}}\{m_0(q+1+qm_0)\}^{1/4}(q+1+qm_0)^{1/2}(3m_0-1)(q+1+qm_0)}{m_0^2(qm_0)^{1/2}} > 0,$$

for small $\epsilon \neq 0$ system (5.9) can be rescaled by $\tilde{u} = d_{23}^{2/5}(\epsilon)u,\ \tilde{v} = d_{23}^{3/5}(\epsilon)v,\ \tilde{\tau} = d_{23}^{-1/5}(\epsilon)\tau$ into the form

$$\dot{\tilde{u}} = \tilde{v}, \qquad \dot{\tilde{v}} = \tilde{G}_1(\tilde{u},\epsilon) + \tilde{G}_2(\tilde{u},\epsilon)\tilde{v} + \tilde{G}_3(\tilde{u},\epsilon)\tilde{v}^2, \tag{5.10}$$

where $\tilde{G}_1(\tilde{u},\epsilon) = d_{23}^{4/5}d_{10}(\epsilon) + \tilde{u}^2 + O(|\tilde{u}|^3)$, $\tilde{G}_2(\tilde{u},\epsilon) = d_{23}^{1/5}d_{20}(\epsilon) + d_{23}^{-1/5}d_{21}(\epsilon)\tilde{u} + d_{23}^{-3/5}d_{22}(\epsilon)\tilde{u}^2 + \tilde{u}^3 + O(|\tilde{u}|^4)$, and $\tilde{G}_3(\tilde{u},\epsilon) = -d_{23}^{-2/5}F_3(-d_{23}^{-2/5}\tilde{u} - c_{11}/(2c_{12}),\epsilon)$, so that the coefficient of the term $\tilde{u}^3\tilde{v}$ in the second equation of (5.10) becomes 1, the same as in the corresponding term in system (5.2). The invertibility of all undergone transformations for small $\epsilon \neq 0$ implies that system (5.10) is topologically conjugate to system (5.2) locally. Hence (5.3) is an induced family of vector fields from system (5.2), the universal unfolding of the degenerate cusp as shown in the Main Theorem in [7]. Comparing the 4-jet of (5.10) with (5.2), we obtain the relation between the induced system and the universal unfolding, i.e.,

$$\mu_1(\epsilon) = d_{23}^{4/5}d_{10}(\epsilon), \quad \mu_2(\epsilon) = d_{23}^{1/5}d_{20}(\epsilon), \quad \mu_3(\epsilon) = d_{23}^{-1/5}d_{21}(\epsilon). \tag{5.11}$$

In particular, $\mu_1(0) = \mu_2(0) = \mu_3(0) = 0$. Computing the Jacobian determinant of relation (5.11) at $(0,0,0)$ with Maple V.7 software, we get

$$\left.\frac{\partial(\mu_1,\mu_2,\mu_3)}{\partial(\epsilon_1,\epsilon_2,\epsilon_3)}\right|_{\epsilon=0} = \frac{1+q+4m_0+3qm_0+5qm_0^2-m_0^2-m_0^3q)(q+1+qm_0)^2}{2},$$

which is > 0 for m_0 near 1. This implies that the induced family (5.3) parameterized by ϵ, and therefore (1.4), is locally equivalent to the unfolding (5.2). The proof is completed. □

Concerning the universal unfolding (5.2), Theorem 4 in [7] gives bifurcation surfaces

$$\begin{array}{lcl}
\mathcal{SN} & = & \{(\mu_1,\mu_2,\mu_3) \in V \,|\, \mu_1 = 0\}, \\
\mathcal{H} & = & \{(\mu_1,\mu_2,\mu_3) \in V \,|\, \mu_2 = \mu_3(-\mu_1)^{\frac{1}{2}} + (-\mu_1)^{\frac{3}{2}} + O((-\mu_1)^{\frac{7}{4}}),\ \mu_1 < 0\}, \\
\mathcal{HL} & =: & \{(\mu_1,\mu_2,\mu_3) \in V \,|\, \mu_2 = \frac{5}{7}\mu_3(-\mu_1)^{\frac{1}{2}} + \frac{103}{77}(-\mu_1)^{\frac{3}{2}} + O((-\mu_1)^{\frac{7}{4}}),\ \mu_1 < 0\}, \\
\mathcal{L} & =: & \{(\mu_1,\mu_2,\mu_3) \in V \,|\, \Xi(\mu_1,\mu_2,\mu_3) = 0,\ \mu_1 < 0\},
\end{array}$$

where V is a neighborhood of $(0,0,0)$ and the surface $\Xi(\mu_1,\mu_2,\mu_3) = 0$ is defined by

$$\mu_2 = (-\mu_1)^{\frac{3}{2}}\left(-\frac{6P(b)}{11P'(b)} + \frac{6b}{11}) + o((-\mu_1)^{\frac{3}{2}}\right), \quad \mu_3 = -\mu_1\left(-\frac{6}{11P'(b)} - \frac{15}{11}\right) + o(-\mu_1)$$

for $\mu_1 < 0$ with the parameter b. Here $P(b)$ is a solution of the Riccati equation $(9b^2-4)P' = 7P^2 + 3bP - 5$, as shown in [7]. Applying the inverse of (5.11) together with (5.1), from $\mathcal{SN}, \mathcal{H}, \mathcal{HL}$, and $\mathcal{L}$ we can give for system (1.4) the corresponding bifurcation surfaces $\mathcal{SN}', \mathcal{H}', \mathcal{HL}'$, and $\mathcal{L}'$, respectively. Thus, Theorem 4 in [7] implies the following.

THEOREM 5.2. *In the (A,p,m)-space there are four surfaces $\mathcal{SN}', \mathcal{H}', \mathcal{HL}'$, $\mathcal{L}'$ near (A_0,p_0,m_0), defined as above, such that system* (1.4) *produces a saddle-node bifurcation near E_0 as (A,p,m) crosses $\mathcal{SN}'$, a Hopf bifurcation near E_0 as (A,p,m)*

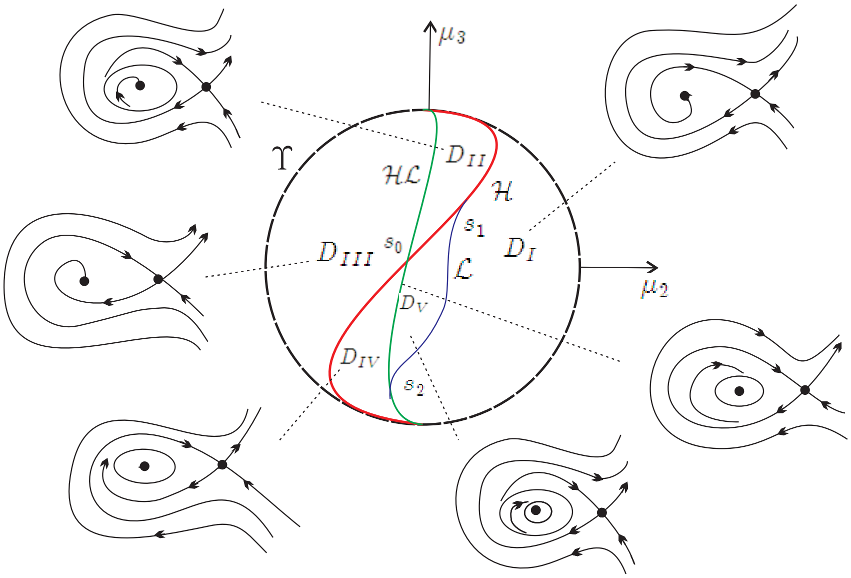


FIG. 3. *Projections of $\mathcal{SN}, \mathcal{H}$, $\mathcal{HL}$, and $\mathcal{L}$ on the (μ_2, μ_3)-plane.*

crosses $\mathcal{H}'$, a homoclinic bifurcation near E_0 as (A, p, m) crosses $\mathcal{HL}'$, and a coalescence of two limit cycles near E_0 as (A, p, m) crosses $\mathcal{L}'$.

The expressions of these bifurcation surfaces given in Theorem 5.2 can be computed by using (5.11) and (5.1). For example, consider $A_0 = 63.73, p_0 = 4.45, m_0 = 1.01$, and $q = 1000$. The surface $\mathcal{HL}'$ is presented as

$$\begin{aligned}&\frac{5}{7}(0.003018260808\epsilon_2 + 1.509170071\epsilon_3)\Theta(\epsilon_1, \epsilon_2, \epsilon_3)\\&+\frac{103}{77}(\Theta(\epsilon_1, \epsilon_2, \epsilon_3))^3 + o(|(\epsilon_1, \epsilon_2, \epsilon_3)|^{3/2}) = 0,\end{aligned}$$

where $\Theta(\epsilon_1, \epsilon_2, \epsilon_3) = \sqrt{0.09569545272\epsilon_1 - 0.003062992329\epsilon_2 - 3.033146367\epsilon_3}$. Parameter values are given by the expression where the homoclinic bifurcation occurs. Expressions of other bifurcation surfaces $\mathcal{SN}', \mathcal{H}'$, and $\mathcal{L}'$ can be given similarly. For a better understanding of the bifurcation diagram, let us observe bifurcation surfaces $\mathcal{SN}, \mathcal{H}, \mathcal{HL}$, and $\mathcal{L}$ in Figure 3. Since each of them is a cone with vertex at the origin (up to a homeomorphism in the parameter space), as in [7], it suffices to observe the bifurcation diagram in a small half ball $S_{\mu_0} = \{(\mu_1, \mu_2, \mu_3) | \, \mu_1^2 + \mu_2^2 + \mu_3^2 < \mu_0, \ \mu_1 < 0\}$ for sufficiently small $\mu_0 > 0$ and project the diagram to the plane $\mathcal{SN}$ (i.e., $\mu_1 = 0$). As indicated in [7], projected on the disk $\Upsilon = S_{\mu_0} \cap \mathcal{SN}$, the curve $\mathcal{H}$ intersects the curve $\mathcal{HL}$ at a point s_0 and $\mathcal{L}$ is tangent to curves $\mathcal{H}$ and $\mathcal{HL}$ at two points s_1 and s_2, respectively. Thus the half ball S_{μ_0} is divided into five open regions D_j $(j = I, II, \dots, V)$, as shown in Figure 3. Let D_j' $(j = I, II, \dots, V)$ be the corresponding regions in the (A, p, m)-space and s_0', s_1', s_2' be the corresponding intersection points, which can be calculated with (5.11). When (A, p, m) lies in these regions, by Theorem 3.1, system (1.4) has two equilibria E_+ and E_- in the interior of the first quadrant and E_- is

TABLE 1
Qualitative properties for various parameters.

(A, p, m)	E_+	Limit cycles	Homoclinic orbits
D'_I	unstable focus or node	no	no
$\mathcal{H}'\backslash\{\widehat{s'_0s'_1}\}$	weak focus(order 1)	no	no
s'_1	weak focus(order 2)	no	no
$\widehat{s'_0s'_1}$	stable weak focus(order 1)	1(order 1)	0
D'_{II}	stable focus or node	1(order 1)	no
$\mathcal{HL}'\backslash\{\widehat{s'_0s'_2}\}$	focus or node	no	1(order 1)
s'_0	stable weak focus(order 1)	no	1(order 1)
D'_{III}	stable focus or node	no	no
D'_{IV}	unstable focus or node	1(order 1)	no
s'_2	unstable focus or node	no	1(order 2)
$\widehat{s'_0s'_2}$	unstable focus or node	1(order 1)	1(order 1)
D'_V	unstable focus or node	2(order 1)	no
$\mathcal{L}'$	unstable focus or node	1(order 2)	0

always a saddle but E_+ is either a focus or a node. Furthermore, by Lemma 4.1 and Theorem 5.2 together with the theorems in [7] and [25], we can list more detailed dynamical behaviors in Table 1, where $\widehat{s'_0s'_1}$ and $\widehat{s'_0s'_2}$ denote the parts of bifurcation surfaces determined by the arcs on $\mathcal{H}$ and $\mathcal{HL}$, respectively, as shown in Figure 3. More concretely, neither a limit cycle nor a homoclinic loop appears in D'_I; a limit cycle arises as parameters go through $\mathcal{H}'$ from D'_I to D'_{II}; the limit cycle expands, deforms into a homoclinic loop and finally breaks as parameters go through $\mathcal{HL}'$ from D'_{II} to D'_{III}; a limit cycle arises again as parameters go through $\mathcal{H}'$ from D'_{III} to D'_{IV}. By continuity, if parameters go through the part of $\mathcal{HL}'$ below s'_2 and return to D'_I from D'_{IV}, the limit cycle disappears; if parameters go from D'_{IV} and hit the arc $\widehat{s'_0s'_2}$, the limit cycle coexists with a homoclinic loop. Furthermore, if parameters enter the region D'_V, the limit cycle persists and another limit cycle arises as the homoclinic loop breaks, i.e., two cycles coexist.

6. Discussion.

The existence of limit cycles in epidemic models can be used to explain oscillatory phenomena observed in the dynamics of some infectious diseases. One of the mechanisms by which epidemic models exhibit periodic oscillations is bifurcation, which occurs when the parameters vary. Early work on studying the dynamics of epidemic models focused on Hopf bifurcation, homoclinic bifurcation, or saddle-node bifurcation separately by using only one bifurcation parameter (Derrick and van den Driessche [6], Hethcote and van den Driessche [14], Liu et al. [17, 18]). Recent studies indicate that some epidemic models undergo codimension 2 bifurcations near degenerate equilibria; i.e., a Bogdanov–Takens bifurcation, which includes a Hopf bifurcation, a homocline bifurcation and a saddle-node bifurcation, can occur when two parameters vary near their critical values (Lizana and Rivero [19], Ruan and Wang [25], Alexander and Moghadas [1, 2], Moghadas [21], Wang [26]). It is interesting to notice that not only epidemic models with nonlinear incidence rates but also simple epidemic models with bilinear mass-action incidence rates can have complex dynamics such as the occurrence of Bogdanov–Takens bifurcations. For instance, Wang and Ruan [27] considered an epidemic model with a bilinear mass-action incidence rate and a constant removal rate of infectious individuals and showed that the model undergoes a sequence of bifurcations, including saddle-node bifurcation, subcritical Hopf bifurcation, and homoclinic bifurcation.

In those epidemic models exhibiting Bogdanov–Takens bifurcations, periodic solutions can arise through a Hopf bifurcation for some parameter values and disappear through a homoclinic bifurcation for some other parameter values, but neither the existence of multiple limit cycles nor the coexistence of a limit cycle and a homoclinic loop is revealed. However, recent work (Alexander and Moghadas [1, 2], Liu, Hethcote, and Levin [17], Moghadas and Alexander [22], Ruan and Wang [25], Wang [26]) indicates that some epidemic models can have two limit cycles. One may expect that the appearance of two limit cycles is due to the fact that degenerate Hopf and degenerate Bogdanov–Takens bifurcations [4] may occur in such epidemic models as well. However, to the best of our knowledge, so far there is no such study on the degenerate Hopf bifurcation and degenerate Bogdanov–Takens bifurcation on epidemic models. One of the difficulties is the lack of general criteria in calculating the multiplicity of a weak focus (see Xiao and Zhu [30] for such a criterion for a predator-prey model; see also Ruan and Xiao [24]).

In this paper, we continued studying the dynamics of a simplified epidemic model (1.3) with a nonlinear incidence rate that was originally considered by Ruan and Wang [25] (see also Liu et al. [17, 18] and Hethcote and van den Driessche [14]). Under certain conditions Ruan and Wang [25] showed that the simplified model (1.3) undergoes a Bogdanov–Takens bifurcation; i.e., it exhibits saddle-node, Hopf, and homoclinic bifurcations. They also established the existence of none, one, or two limit cycles. In this paper, we first calculated the second order Liapunov value of the weak focus and proved that the maximal multiplicity of the weak focus is 2 by technically dealing with some complicated multivariable polynomials, which implies that at most two limit cycles can arise near the weak focus. Then, by reducing the determination of the sign for polynomials of higher degrees to revised sign lists, we re-established the uniqueness of the limit cycle. Finally, we reduced system (1.4) to a form of universal unfolding for a cusp of codimension 3 and showed the coexistence of limit cycles and homoclinic loops via a degenerate Bogdanov–Takens bifurcation.

The coexistence of limit cycles and homoclinic loops demonstrates that epidemic models with saturated incidence rates exhibit very different and complex dynamics. Furthermore, the results indicate that the dynamical behavior of the model is very sensitive to the initial densities of the susceptible and infectious individuals. When the initial values lie inside the homoclinic loop, the numbers of susceptible and infectious individuals fluctuate periodically about the endemic levels. Such periodic patterns will be helpful in designing control and intervention policies for the disease. When the initial values lie outside the homoclinic loop, the disease will die out even if there are two endemic equilibria (see Figure 3). This means that the disease can be controlled and eradicated even above the threshold.

To the best of our knowledge, this is the first time that a limit cycle and a homoclinic loop have been shown to coexist in a realistic epidemic model. Though we focused on a simple case of SIRS models with a specific saturated incidence rate, we believe that such rich and complex dynamics can occur in other epidemic models with general saturated incidence rates as well as other types of nonlinear incidence rates (Hethcote and van den Driessche [14], Liu et al. [17, 18]).

Appendix.

(A1) As claimed in section 4, for each $j = 1, \dots, 13$, the polynomial ϖ_j in variables

$\alpha_1, \ldots, \alpha_6$, being the jth principal minor of the matrix

$$\mathrm{Discr}(\Psi) = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6\alpha_0 & 5\alpha_1 & 4\alpha_2 & 3\alpha_3 & 2\alpha_4 & \alpha_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 \end{bmatrix},$$

can be calculated directly with Maple software. For example, $\varpi_1 = \alpha_0, \varpi_2 = 6\alpha_0^2, \varpi_3 = \alpha_0^2\alpha_1, \varpi_4 = -\alpha_0^2(-5\alpha_1^2 + 12\alpha_2\alpha_0)$, $\varpi_5 = \alpha_0^2(\alpha_1^2\alpha_2 - 4\alpha_0\alpha_2^2 + 3\alpha_0\alpha_3\alpha_1)$, $\varpi_6 = 2\alpha_0^2(24\alpha_0^2\alpha_4\alpha_2 - 27\alpha_0^2\alpha_3^2 - 8\alpha_0\alpha_2^3 + 24\alpha_0\alpha_3\alpha_1\alpha_2 - 10\alpha_0\alpha_4\alpha_1^2 + 2\alpha_2^2\alpha_1^2 - 5\alpha_3\alpha_1^3)$. The expressions for others will be much longer.

(A2) The polynomials $\vartheta_j, \iota_{ij}, \omega_j$ claimed in (5.4), a_{ij} and b_{ij} in (5.6), c_{ij} and F_3 in (5.7), and d_{ij} in (5.9) can be calculated directly with Maple software. Except for

$$c_{10} = a_{20}b_{10} - a_{10}b_{20},\ c_{11} = a_{20}b_{11} - a_{10}b_{21} + a_{21}b_{10} - a_{11}b_{20},\ c_{12} = -\sqrt{2m_0(q+1+qm_0)}(q+1+qm_0)/(4qm_0) + O(|\epsilon|),$$
$$c_{13} = -b_{22}a_{11} + b_{13}a_{20} - b_{23}a_{10} + b_{10}a_{23} - b_{21}a_{12} + b_{12}a_{21} - b_{20}a_{13} + b_{11}a_{22},$$
$$c_{20} = (a_{11}a_{20} + b_{20}a_{20} - a_{10}a_{21})/a_{20}, \quad c_{21} = -(2a_{10}a_{22}a_{20} - bb21a_{20}^2 - 2a_{12}a_{20}^2 + a_{21}a_{11}a_{20} - a_{10}a_{21}^2)/a_{20}^2,$$
$$c_{22} = -(-3a_{13}a_{20}^3 + 2a_{22}a_{20}^2a_{11} + a_{21}a_{12}a_{20}^2 + 3a_{10}a_{23}a_{20}^2 - b_{22}a_{20}^3 - 3a_{22}a_{20}a_{10}a_{21} - a_{21}^2a_{11}a_{20} + a_{10}a_{21}^3)/a_{20}^3,$$
$$c_{23} = -\{\sqrt{2}(3m_0 - 1)(q + 1 + qm_0)^3\}/\{4m_0^2q\sqrt{m_0(q + 1 + qm_0)}\} + O(|\epsilon|),$$
$$d_{10} = (4c_{10}c_{12} - c_{11}^2)/(4c_{12}^2) + O(|\epsilon|^2), \quad d_{20} = -(-8c_{20}c_{12}^3 + 4c_{21}c_{11}c_{12}^2 + c_{23}c_{11}^3 - 2c_{22}c_{11}^2c_{12})/(8c_{12}^3\sqrt{-c_{12}}) + O(|\epsilon|^2),$$
$$d_{21} = -(-4c_{22}c_{11}c_{12} + 4c_{21}c_{12}^2 + 3c_{23}c_{11}^2)/(4c_{12}^2\sqrt{-c_{12}}) + O(|\epsilon|^2), \quad d_{22} = (2c_{22}c_{12} - 3c_{23}c_{11})/(2c_{12}\sqrt{-c_{12}}) + O(|\epsilon|^2),$$
$$d_{23} = (\frac{1}{2})^{3/4}(\sqrt{m_0(q + 1 + qm_0)}(q + 1 + qm_0)/(qm_0))^{1/2}(3m_0 - 1)(q + 1 + qm_0)/m_0^2 + O(|\epsilon|),$$
$$F_3(\tilde{X}) = a_{21}/a_{20} + \{(-a_{21}^2 + 2a_{22}a_{20})/a_{20}^2\}\tilde{X} - \{(3a_{21}a_{22}a_{20} - 3a_{23}a_{20}^2 - a_{21}^3)/a_{20}^3\}\tilde{X}^2 - \{(4a_{21}a_{23}a_{20}^2 + 2a_{22}^2a_{20}^2 - 4a_{22}a_{20}a_{21}^2 + a_{21}^4)/a_{20}^4\}\tilde{X}^3 + O(|\tilde{X}^4|).$$

The other polynomials have long expressions. Their presentations and the Maple scripts are available upon request.

Acknowledgment. The authors are very grateful to the referees for their helpful comments and suggestions.

REFERENCES

[1] M. E. Alexander and S. M. Moghadas, *Periodicity in an epidemic model with a generalized non-linear incidence*, Math. Biosci., 189 (2004), pp. 75–96.

[2] M. E. Alexander and S. M. Moghadas, *Bifurcation analysis of an SIRS epidemic model with generalized incidence*, SIAM J. Appl. Math., 65 (2005), pp. 1794–1816.

[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, UK, 1992.

[4] S. M. Baer, B. W. Kooi, Y. A. Kuznetsov, and H. R. Thieme, *Multiparametric bifurcation analysis of a basic two-stage population model*, SIAM J. Appl. Math., 66 (2006), pp. 1339–1365.

[5] V. Capasso and G. Serio, *A generalization of the Kermack–Mckendrick deterministic epidemic model*, Math. Biosci., 42 (1978), pp. 43–61.

[6] W. R. Derrick and P. van den Driessche, *Homoclinic orbits in a disease transmission model with nonlinear incidence and nonconstant population*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 299–309.

[7] F. Dumortier, R. Roussarie, and J. Sotomayor, *Generic* 3*-parameter families of vector fields on the plane, unfolding a singularity with nilpotent linear part. The cusp case of codimension* 3, Ergodic Theory Dynam. Systems, 7 (1987), pp. 375–413.

[8] M. G. M. Gomes, A. Margheri, G. F. Medley, and C. Rebelo, *Dynamical behaviour of epidemiological models with sub-optimal immunity and nonlinear incidence*, J. Math. Biol., 51 (2005), pp. 414–430.

[9] J. K. Hale, *Ordinary Differential Equations*, 2nd ed., Wiley-Interscience, New York, 1980.

[10] H. W. Hethcote, *A thousand and one epidemic models*, in Frontiers in Theoretical Biology, S. A. Levin, ed., Lecture Notes in Biomath. 100, Springer-Verlag, Berlin, 1994, pp. 504–515.

[11] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.

[12] H. W. Hethcote and S. A. Levin, *Periodicity in epidemiological models*, in Applied Mathematical Biology, Biomath. Texts 18, S. A. Levin, T. G. Hallam and L. J. Gross, eds., Springer-Verlag, New York, 1989, pp. 193–211.

[13] H. W. Hethcote, H. W. Stech, and P. van den Driessche, *Periodicity and stability in epidemic models: A survey*, in Differential Equations and Applications in Ecology, Epidemics and Population Problems, S. N. Busenberg and K. L. Cook, eds., Academic Press, New York, 1981, pp. 65–82.

[14] H. W. Hethcote and P. van den Driessche, *Some epidemiological models with nonlinear incidence*, J. Math. Biol., 29 (1991), pp. 271–287.

[15] R. E. Kooij and A. Zegeling, *A predator-prey model with Ivlev's functional response*, J. Math. Anal. Appl., 198 (1996), pp. 473–489.

[16] C.-Z. Li and C. Rousseau, *A system with three cycles appearing in a Hopf bifurcation and dying in a homoclinic bifurcation: The cusp of order* 4, J. Differential Equations, 23 (1986), pp. 187–204.

[17] W. Liu, H. W. Hethcote, and S. A. Levin, *Dynamical behavior of epidemiological models with non-linear incidence rate*, J. Math. Biol., 25 (1987), pp. 359–380.

[18] W. Liu, S. A. Levin, and Y. Iwasa, *Influence of nonlinear incidence rate upon the behavior of SIRS epidemiological models*, J. Math. Biol., 23 (1986), pp. 187–204.

[19] M. Lizana and J. Rivero, *Multiparametric bifurcations for a model in epidemiology*, J. Math. Biol., 35 (1996), pp. 21–36.

[20] N. G. Lloyd, *Limit cycles of polynomial systems—some recent developments*, in New Directions in Dynamical Systems, T. Bedford and J. Swift, eds., LMS Lect. Notes 127, Cambridge University Press, Cambridge, UK, 1988, pp. 192–238.

[21] S. M. Moghadas, *Analysis of an epidemic model with bistable equilibria using the Poincaré index*, Appl. Math. Comput., 149 (2004), pp. 689–702.

[22] S. M. Moghadas and M. E. Alexander, *Bifurcations of an epidemic model with non-linear incidence and infection-dependent removal rate*, Math. Med. Biol., 23 (2006), pp. 231–254.

[23] R. R. Regoes, D. Ebert, and S. Bonhoeffer, *Dose-dependent infection rates of parasites produce the Allee effect in epidemiology*, Proc. Roy. Soc. London Ser. B, 269 (2002), pp. 271–279.

[24] S. Ruan and D. Xiao, *Global analysis in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 61 (2001), pp. 1445–1472.

[25] S. Ruan and W. Wang, *Dynamical behaviour of an epidemic model with a nonlinear incidence rate*, J. Differential Equations, 188 (2003), pp. 135–163.

[26] W. Wang, *Epidemic models with nonlinear infection forces*, Math. Biosci. Engrg., 3 (2006), pp. 267–279.

[27] W. Wang and S. Ruan, *Bifurcations in an epidemic model with constant removal rate of the infectives*, J. Math. Anal. Appl., 291 (2004), pp. 775–793.

[28] D. Xiao and S. Ruan, *Global analysis of an epidemic model with nonmonotone incidence rate*, Math. Biosci., 208 (2007), pp. 419–429.

[29] D. Xiao and Z.-F. Zhang, *On the uniqueness and nonexistence of limit cycles for predator-prey systems*, Nonlinearity, 16 (2003), pp. 1185–1201.

[30] D. Xiao and H. Zhu, *Multiple focus and Hopf bifurcations in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 66 (2006), pp. 802–819.

[31] L. Yang, *Recent advances on determining the number of real roots of parametric polynomials*, J. Symbolic Comput. 28 (1999), pp. 225–242.

[32] Z.-F. Zhang, T.-R. Ding, W.-Z. Huang, and Z.-X. Dong, *Qualitative Theory of Differential Equations*, AMS, Providence, RI, 1992.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 641–669

PHASE FIELD MODELS FOR BIOFILMS. I. THEORY AND ONE-DIMENSIONAL SIMULATIONS*

TIANYU ZHANG†, N. G. COGAN†, and QI WANG†

Abstract. We derive a set of phase field models for biofilms using the one-fluid two-component formulation in which the combination of extracellular polymeric substances (EPS, or polymer networks) and the bacteria is effectively modeled as one fluid component, while the collective ensemble of nutrient substrates and the solvent are modeled as the other. The biofilm is assumed to be an incompressible continuum, in which the motion of the polymer network and the solvent relative to the average velocity is accounted for by binary mixing kinetics. Various constitutive stress models are proposed for the effective polymer network component according to the property of the polymer network. Steady states are identified, their stability is analyzed (where two long wave growth modes are identified), and numerical solutions of different variations of the model in one space dimension are discussed and compared.

Key words. biofilms, Cahn–Hilliard equation, finite difference scheme, phase field, polymer networks, steady states, stability

AMS subject classifications. 92C37, 65M06, 92C15

DOI. 10.1137/070691966

1. Introduction.

Biofilms are ubiquitous in natural and industrial settings. They exist on wet surfaces and consist of myriad microbes, their byproducts, and trapped particles. A biofilm community can be formed by a single bacterial species, but in nature biofilms almost always consist of rich mixtures of many species of bacteria, as well as fungi, algae, yeasts, protozoa, other microorganisms, debris, and corrosion products. Biofilms are held together primarily by polysaccharides and other long chain molecules, collectively termed "extracellular polymeric substances" or EPS. The bacteria cells produce the EPS and are held together by EPS strands, allowing them to develop complex, three-dimensional, resilient, attached communities [7, 9, 11, 12, 17, 18, 19, 24].

The Center for Disease Control and National Institutes of Health recently estimated that 65% to 85% of all chronic infections can be attributed to bacterial biofilms [10]. In human diseases, biofilm infections are some of the most difficult to treat. Even with rigorous antibiotic regimens, some biofilms, such as those within the thick airway mucus of cystic fibrosis (CF) patients, persist throughout the course of the disease process [16]. Bacterial biofilms can also be utilized in bioterrorism in which persistent "bioterrorist agent biofilms" formed by *Francisella tularensis* can grow on surfaces where environmental amoebas can phagocytose them, allowing for growth of fibrosis [16].

Biofilms cost the U.S. literally billions of dollars every year in energy losses, equipment damage, product contamination, and medical infections. Understanding the dynamics of the growth, transport, and destruction of biofilms is important for improving water treatment and medical treatment of diseases, protecting equipment

*Received by the editors May 15, 2007; accepted for publication (in revised form) April 15, 2008; published electronically December 5, 2008. This work was supported by the Air Force Office of Scientific Research, Air Force Materials Command, USAF, under grant number F49550-05-1-0025 and the National Science Foundation through grants DMS-0605029, DMS-0626180, and DMS-0548511.
http://www.siam.org/journals/siap/69-3/69196.html

†Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510 (zhang@math.fsu.edu, cogan@math.fsu.edu, wang@math.fsu.edu).

or devices from corrosion, and even preventing bioterrorism. The improved understanding of biofilms will have a significant impact on environmental sciences, medicine, civil engineering, naval sciences, military applications, and homeland security.

There have been increasing efforts to model biofilm structures and dynamics over the last two decades [25, 26, 27, 28], in which methods based on cellular automata, particle-based methods, continuum models, and multispecies modeling are attempted [1, 13, 20, 21, 30]. Recently Cogan and Keener developed a two-fluid model for biofilms, treating bacteria as a part of the polymer network [7]. The nutrient substrate is also treated passively as a part of the solvent. This work extended the polymeric mixture models of Tanaka [29] and Milner [23] and the work of Wolgemuth et al. [31] for biological material mixtures. Similar multifluid modeling extension has also been done by Klapper and colleagues [1, 20].

We briefly recall the multifluid theory of Cogan and Keener for biofilms next. Let ϕ_n be the volume fraction of the polymer network, ϕ_s that of the solvent, $\mathbf{v}_n$ the velocity of the polymer network, $\mathbf{v}_s$ the velocity of the solvent, c the concentration of the nutrient substrate, p the pressure, and τ_n and τ_s the network and solvent stress tensor, respectively. In the Cogan–Keener model, the substrate is passively treated as a part of the solvent. This two-fluid model consists of the linear momentum balance equation for each fluid, where inertia for all species are ignored, and the transport equation for the nutrient concentration as well as the volume fraction of the polymer network [7].

The momentum balance equation for each species is respectively given by

$$\begin{gathered}\nabla\cdot(\phi_n\tau_n) - h_f\phi_n\phi_s(\mathbf{v}_n-\mathbf{v}_s) - \nabla\Psi - \phi_n\nabla p = 0,\\ \nabla\cdot(\phi_s\tau_s) + h_f\phi_n\phi_s(\mathbf{v}_n-\mathbf{v}_s) - \phi_s\nabla p = 0,\end{gathered}\tag{1.1}$$

where h_f is the coefficient of friction and Ψ is the osmotic pressure due to the existence of the polymer network; the transport equation of the polymer volume fraction and the conservation of the volume fraction for the solvent are respectively given by

$$\begin{gathered}\frac{\partial\phi_n}{\partial t} + \nabla\cdot(\phi_n\mathbf{v}_n) = g_n,\\ \frac{\partial\phi_s}{\partial t} + \nabla\cdot(\phi_s\mathbf{v}_s) = 0,\end{gathered}\tag{1.2}$$

and the equation for the nutrient substrate consumption is given by

$$\frac{\partial}{\partial t}(\phi_s c) + \nabla\cdot(c\mathbf{v}_s\phi_s - D_s\phi_s\nabla c) = -g_c,\tag{1.3}$$

where g_n is the production rate for the polymer network, g_c is the consumption rate of the nutrient substrate in the solvent, and D_s is the diffusion constant of the nutrient substrate.

Both the polymer network and the solvent are assumed viscous in the Cogan–Keener model. The extra stress tensor, the osmotic pressure, and the production as well as consumption rates are given by the following constitutive laws:

$$\begin{aligned}&\tau_n = 2\eta_n\mathbf{D}_n,\\ &\tau_s = 2\eta_s\mathbf{D}_s,\\ &g_c = \phi_n A c,\\ &g_n = \epsilon\mu\phi_n\frac{c}{K_c+c},\\ &\Psi = \frac{kT}{v_1}\left[\ln(1-\phi_n) + \left(1-\frac{1}{N}\right)\phi_n + \chi\phi_n^2\right],\end{aligned}\tag{1.4}$$

where $\eta_{n,s}$ are the viscosity of the network and the solvent, respectively, $\mathbf{D}_{n,s} = \frac{1}{2}[\nabla \mathbf{v}_{n,s} + \nabla \mathbf{v}_{n,s}^T]$ is the rate of strain tensor for the network and the solvent, respectively, A is the consumption rate of the substrate, μ is the maximum production rate, K_c is the half-saturation constant, ϵ is a scaling parameter, N is the polymerization index, v_1 is the volume of the solvent molecule, k is the Boltzmann constant, T is the temperature, and χ is the Flory–Huggins mixing parameter [14, 15]. We note that the equation for the concentration of bacteria is a decoupled equation in the Cogan–Keener model and is therefore not listed above.

Given that

$$\phi_n + \phi_s = 1, \tag{1.5}$$

the following constraints arise:

$$\begin{aligned} \nabla \cdot (\phi_n \mathbf{v}_n + \phi_s \mathbf{v}_s) = g_n, \\ \nabla \cdot (\phi_n \tau_n + \phi_s \tau_s) = \nabla(\Psi + p). \end{aligned} \tag{1.6}$$

We note that $\mathbf{v} = \phi_n \mathbf{v}_n + \phi_s \mathbf{v}_s$ is the volume averaged velocity. Clearly, it is not divergence-free when $g_n \neq 0$, indicating that the material is in fact “compressible.” The second constraint gives the force balance equation for the volume averaged stress.

We note that the constraint above leads the bulk volume of the two-fluid material system to increase when $g_n \neq 0$. Practically, the individual velocity of each species is hardly measurable; moreover, it is impossible to impose the boundary conditions for velocities at inflow and outflow boundaries for each species. Therefore two-fluid theories are not easy to adopt in fluid dynamics and rheological studies. The practical use of the two-fluid models includes ignoring the solvent velocity [7], ignoring the stress deformation [20], or simply imposing periodic boundary conditions. This clearly limits the applicability of the multifluid biofilm theories.

In this paper, we embark on a different approach, assuming the biofilm-solvent mixture is incompressible, whose bulk motion is measured by a divergence-free averaged velocity field, adopting the one-fluid multicomponent formulation for mixture theories [2]. We retain the effective treatment of the polymer network/bacteria and substrate/solvent combinations. The excessive velocity in addition to the average one is accounted for by polymer-solvent mixing dynamics. Through an essentially mean field approach, we can couple the polymer network deformation and biofilm/solvent interfacial dynamics into the fluid mixture motion, which to the best of our knowledge has not been done to biofilm models systematically so far. The effective polymer comprising the EPS and bacteria is modeled as a viscoelastic “solution” in which the bacterium is the solution since it is viscous while the EPS is modeled as a linear polymer strand of a network [3].

The rest of this paper is organized as follows. First we develop a set of phase field models for biofilms by accounting for the transport of polymer networks, nutrient substrates, and the response of the polymer network in flow in several plausible ways within the theoretical framework of one-fluid multicomponent systems. We then analyze the stability of some steady states to investigate possibly unstable modes. Finally, we numerically study the biofilm growth and expansion in one space dimension and compare the results with respect to various formulations of the mixture theory.

2. Mathematical models. We study the biofilm in solvent as a fluid mixture of two components: the effective polymer network encompassing the bacteria trapped inside and the effective solvent which includes the nutrient substrates and pure solvent.

We adopt the one-fluid two-component formalism for fluid mixtures to develop a single-fluid, multicomponent model using the volume averaged velocity and the volume fractions of the two distinctive components. The polymer network volume fraction ϕ_n plays the role of a phase field variable in the theory. When $\phi_n = 0$, the fluid consists of entirely the solvent; otherwise, it is a true binary mixture when $0 < \phi_n < 1$. (The case of $\phi_n = 1$ is excluded in biofilms since they are never dry.) Therefore, the resulting theory is an effective phase field model. The two distinctive phases are modeled by $\phi_n = 0$ and $\phi_n > 0$, respectively. The inhomogeneity of the biofilm is accounted for by the variation of ϕ_n.

2.1. Phase field formulation. When the fluid mixture is incompressible, the average velocity is divergence-free. The bulk fluid is convected by the average velocity. In addition to the bulk convection, the polymer network is also transported by an additional flux due to mixing of two different components. Specifically, the local instantaneous flux consists of two parts: the flux convected by the average velocity $\mathbf{v}$ and the excessive flux due to polymer-solvent binary mixing. The latter contribution to the flux of the polymer volume fraction is assumed proportional to the mixing force given by the gradient of the free energy variation

$$\mathbf{f}_n = -\lambda_{ch} \nabla \frac{\delta f}{\delta \phi_n}, \tag{2.1}$$

where λ_{ch} is the proportionality parameter that has the same unit as the mobility. This is consistent with the Ginzburg–Landau dynamics in condensed matter physics [6]. The mixing free energy density f as a function of ϕ_n is given by the extended Flory–Huggins free energy density [14, 15]

$$f = kT \left[\frac{\gamma_1}{2} \|\nabla \phi_n\|^2 + \gamma_2 \left(\frac{\phi_n}{N} \ln \phi_n + (1-\phi_n)\ln(1-\phi_n) + \chi \phi_n (1-\phi_n) \right) \right], \tag{2.2}$$

where γ_1 and γ_2 measure the strength of the distortional and bulk mixing free energy, respectively, χ is the Flory–Huggins mixing parameter, N is the generalized polymerization index, $1/\gamma_2$ is proportional to the specific volume of the solvent molecule, and $\| \cdot \|$ denotes the l_2 norm of a vector in $\mathbf{R}^3$. The distortional free energy is included in the extended Flory–Huggins mixing free energy to account for the surface tension effect at the solvent-biofilm interface defined by $\{\mathbf{x} | \phi_n(\mathbf{x}, t) = \epsilon \text{ as } \epsilon \to 0^+\}$ and penalizing spatial inhomogeneity in the mixture. The variation of f with respect to ϕ_n (known as the chemical potential) is given by

$$\frac{\delta f}{\delta \phi_n} = -kT \left[\gamma_1 \Delta \phi_n + \gamma_2 \left[-\frac{1}{N} - \frac{\ln \phi_n}{N} + \ln(1-\phi_n) + 1 - \chi + 2\chi\phi_n \right] \right]. \tag{2.3}$$

Representing the growth rate of the polymer network produced by bacteria as the reaction rate for the polymer volume fraction, we propose the transport equation for the volume fraction of the polymer network as follows:

$$\frac{\partial \phi_n}{\partial t} + \nabla \cdot (\phi_n \mathbf{v}) = \nabla \cdot \left(\lambda_{ch} \nabla \frac{\delta f}{\delta \phi_n} \right) + g_n. \tag{2.4}$$

This is the Cahn–Hilliard equation [4, 5] with a reaction term (polymer production). From the given excessive flux, we can identify the instantaneous excessive velocity as

$$\mathbf{v}_n^e = -\lambda_{ch} \frac{1}{\phi_n} \nabla \frac{\delta f}{\delta \phi_n} \tag{2.5}$$

when $\phi_n \neq 0$. It is zero when $\phi_n = 0$.

Another form of the transport equation for ϕ_n can be obtained by arguing that the excessive flux is due to an excessive velocity which is proportional to the mixing force and takes the form $\mathbf{v}_n^e = -\lambda \nabla \frac{\delta f}{\delta \phi_n}$, in which the excessive flux is given by $-\lambda \phi_n \nabla \frac{\delta f}{\delta \phi_n}$. Here λ is the mobility parameter. This can also be obtained from the Ginzburg–Landau dynamics by assuming that λ_{ch} is proportional to the polymer volume fraction: $\lambda_{ch} = \lambda \phi_n$. The transport equation for ϕ_n is given by

$$\frac{\partial \phi_n}{\partial t} + \nabla \cdot (\phi_n \mathbf{v}) = \nabla \cdot \left[\lambda \phi_n \nabla \frac{\delta f}{\delta \phi_n} \right] + g_n. \tag{2.6}$$

This is called the modified or singular Cahn–Hilliard equation. When the fluid is entirely occupied by polymer networks, one of the extreme cases, we argue that mixing will cease. Therefore, it is plausible to assume that the mobility matrix is proportional to the solvent volume fraction as well:

$$\lambda = \lambda_0 (1 - \phi_n). \tag{2.7}$$

However, this perhaps would never happen in biofilm materials since biofilms always contain solvent in their sponge-like structures. Both the Cahn–Hilliard and the modified Cahn–Hilliard models will be tested in the following. The numerical simulation presented in later sections shows that the modified Cahn–Hilliard equation is more appropriate for the transport of ϕ_n, especially with the polymer production included in the transport.

The remaining governing equations for the mixture consist of the continuity equation, the momentum transport or balance equation, and the transport equation for the nutrient:

$$\begin{aligned} &\nabla \cdot \mathbf{v} = 0, \\ &\rho \frac{d\mathbf{v}}{dt} = \nabla \cdot (\tau_{extra}) - [\nabla p + \gamma_1 kT \nabla \cdot (\nabla \phi_n \nabla \phi_n)], \\ &\frac{\partial}{\partial t}(\phi_s c) + \nabla \cdot (c \mathbf{v} \phi_s - D_s \phi_s \nabla c) = -g_c, \end{aligned} \tag{2.8}$$

where ρ_s and ρ_n are the density of the solvent and polymer, respectively, $\rho = \phi_s \rho_s + \phi_n \rho_n$ is the averaged density, and τ_{extra} is the total extra bulk stress for the mixture. Here g_n, g_c are the reaction rates defined in (1.4). We note that when the densities of the polymer network and solvent are equal, the density of the mixture is a constant and the volume fraction averaged velocity is the mass averaged velocity.

In the above momentum balance equation, the presence of the extra term $\gamma_1 kT \nabla \cdot (\nabla \phi_n \nabla \phi_n)$ is due to the spatial inhomogeneity resulting from a virtual work principle [22]. The nutrient transport is assumed to be convected by the average velocity. The incompressibility condition $\nabla \cdot \mathbf{v} = 0$ and the constraint $\phi_n + \phi_s = 1$ require that the transport equation for ϕ_s have a decay term $-g_n$, leading to

$$\frac{\partial \phi_s}{\partial t} + \nabla \cdot (\phi_s \mathbf{v}) = -\nabla \cdot \left(\lambda_{ch} \nabla \frac{\delta f}{\delta \phi_n} \right) - g_n \tag{2.9}$$

in the Cahn–Hilliard model or

$$\frac{\partial \phi_s}{\partial t} + \nabla \cdot (\phi_s \mathbf{v}) = -\nabla \cdot \lambda \left(\phi_n \nabla \frac{\delta f}{\delta \phi_n} \right) - g_n \tag{2.10}$$

in the modified Cahn–Hilliard model. In the Cahn–Hilliard model, the excessive solvent velocity can be identified as

$$\mathbf{v}_s^e = \lambda_{ch}\frac{1}{\phi_s}\nabla\frac{\delta f}{\delta\phi_n}, \tag{2.11}$$

whereas the velocity is given by

$$\mathbf{v}_s^e = \lambda\frac{\phi_n}{\phi_s}\nabla\frac{\delta f}{\delta\phi_n} \tag{2.12}$$

in the modified Cahn–Hilliard model. The actual solvent velocity can be calculated by

$$\mathbf{v}_s = \mathbf{v} + \mathbf{v}_s^e. \tag{2.13}$$

Analogously, the polymer network velocity is given by

$$\mathbf{v}_n = \mathbf{v} + \mathbf{v}_n^e. \tag{2.14}$$

With this definition, we easily see that the average velocity is indeed the volume averaged velocity

$$\mathbf{v} = \phi_n\mathbf{v}_n + \phi_s\mathbf{v}_s. \tag{2.15}$$

In the above formulation of the theory, the nutrient substrate is assumed to be transported along with the average velocity. If we assume that the nutrient is transported with the solvent velocity instead, the nutrient transport equation is given by

$$\frac{\partial}{\partial t}(\phi_s c) + \nabla\cdot(c\mathbf{v}_s\phi_s - D_s\phi_s\nabla c) = -g_c. \tag{2.16}$$

2.2. Constitutive equations for effective polymer. The extra stress for the polymer network–solvent mixture will supply the crucial link to complete the governing system of equations for the biofilm model. The simplest choice is treating the polymer-solvent mixture as an extended Newtonian fluid like in (2.17). When both the solvent and the polymer are modeled as viscous fluids, the constitutive equations for the extra stresses are given by

$$\tau_n = 2\eta_n\mathbf{D}, \qquad \tau_s = 2\eta_s\mathbf{D}, \tag{2.17}$$

where $\mathbf{D} = \frac{1}{2}[\nabla\mathbf{v} + \nabla\mathbf{v}^T]$ is the rate of strain tensor and η_n, η_s are the polymer and solvent viscosities, respectively. Alternatively, we assume the extra stress to be proportional to the rate of strain tensor given by the velocity field of each component:

$$\tau_n = 2\eta_n\mathbf{D}_n, \qquad \tau_s = 2\eta_s\mathbf{D}_s, \tag{2.18}$$

where $\mathbf{D}_n = \frac{1}{2}(\nabla\mathbf{v}_n + \nabla\mathbf{v}_n^T)$, $\mathbf{D}_s = \frac{1}{2}(\nabla\mathbf{v}_s + \nabla\mathbf{v}_s^T)$. To account for the shear thinning effect, the polymer viscosity could depend on the rate of strain tensor like the power-law type [3].

However, because biofilms are hydrogels, they exhibit elastic and/or viscoelastic behavior depending on the time-scale of interest. To account for these contributions of the network, more sophisticated constitutive equations should be employed. We propose both an elastic and a viscoelastic model next. Given the composition of the effective polymer network, the stress associated to it should contain a viscous part accounting for the stress due to the viscous bacterial component denoted by τ_{ns}. It has two variations

$$\tau_{ns} = 2\eta_n \mathbf{D} \quad \text{or} \quad \tau_{ns} = 2\eta_n \mathbf{D}_n, \tag{2.19}$$

where η_n is the bacterial contribution to the polymeric viscosity in the effective polymer.

Rubber-elastic model. We model the EPS network as a gel. According to rubber-elastic theory, the elastic constitutive equation is given by

$$\tau_n = \nu kT \mathbf{F} \cdot \mathbf{F}^T = \nu kT \mathbf{B}, \tag{2.20}$$

where $\mathbf{F}$ is the deformation gradient tensor, $\mathbf{B} = \mathbf{F} \cdot \mathbf{F}^T$ is the Finger tensor, and v is the polymer number density. The time evolution of the deformation gradient tensor in the absence of solvent is given by

$$\frac{d\mathbf{F}}{dt} = \nabla \mathbf{v}_n \cdot \mathbf{F}, \tag{2.21}$$

where $\mathbf{v}_n$ is the polymer network velocity. The time evolution of the elastic stress tensor (as well as Finger tensor $\mathbf{B}$) follows the equation

$$\frac{\partial \tau_n}{\partial t} + \mathbf{v}_n \cdot \nabla(\tau_n) - [\nabla \mathbf{v}_n \cdot \tau_n + \tau_n \cdot \nabla \mathbf{v}_n^T] = 0. \tag{2.22}$$

An alternative choice for the rate-of-strain tensor is the rate of strain associated with the average velocity. Then, the constitutive equation for the elastic stress tensor is given by

$$\frac{d\tau_n}{dt} - [\nabla \mathbf{v} \cdot \tau_n + \tau_n \cdot \nabla \mathbf{v}^T] = 0, \tag{2.23}$$

where $\frac{d}{dt}(\bullet) = \frac{\partial}{\partial t}(\bullet) + \mathbf{v} \cdot \nabla(\bullet)$ is the material derivative and the polymer network is assumed to deform with the average velocity gradient.

Johnson–Segalman model. Considering the creation and annihilation rate for the network strands or segments in the network, we adopt the temporary network model for the viscoelastic EPS [3]. When the two rates are balanced, the constitutive equation for the elastic stress tensor is given by the following Johnson–Segalman model:

$$\frac{\partial \tau_n}{\partial t} + \nu_n \nabla \cdot (\mathbf{v}_n \tau_n) - \mathbf{W}_n \cdot \tau_n + \tau_n \cdot \mathbf{W}_n - a[\mathbf{D}_n \cdot \tau_n + \tau_n \cdot \mathbf{D}_n] + \frac{\tau_n}{\lambda_1} = \frac{2\eta_p}{\lambda_1} \mathbf{D}_n, \tag{2.24}$$

where a is a rate parameter between -1 and 1, λ_1 is the EPS relaxation time, and η_p is the EPS polymer network viscosity in the effective polymer [3]. $a = 1$ yields the Oldroyd-B model with the upper convected derivative, and $a = -1$ corresponds to the lower convected derivative. The rubber-elastic model can be viewed as a limiting case of the current model as $\lambda_1 \to \infty$ and $a = 1$; the viscous limit is recovered if $\lambda_1 \to 0$; whereas the highly elastic model is the limit of $\lambda_1 \to \infty$, $\frac{\eta_p}{\lambda_1} \to G$, where G is the elastic modulus.

An alternative formulation is to replace the network velocity $\mathbf{v}_n$ by the average velocity $\mathbf{v}$ analogous to the rubber-elastic case. The constitutive equation for the extra stress is then given by

$$\frac{d\tau_n}{dt} - \mathbf{W}\cdot\tau_n + \tau_n\cdot\mathbf{W} - a[\mathbf{D}\cdot\tau_n + \tau_n\cdot\mathbf{D}] + \frac{\tau_n}{\lambda_1} = \frac{2\eta_p}{\lambda_1}\mathbf{D}. \tag{2.25}$$

In summary, the phase field theories for biofilms consist of four sets of equations of multiple variations. In the following, the suffix A indicates that the average velocity is used, while N denotes that the network and the solvent velocity, respectively, are used.

Momentum and continuity equation.

$$\begin{aligned} &\nabla\cdot\mathbf{v} = 0,\\ &\rho\frac{d\mathbf{v}}{dt} = \nabla\cdot(\tau_{extra}) - [\nabla p + \gamma_1 kT\nabla\cdot(\nabla\phi_n\nabla\phi_n)],\\ &\tau_{extra} = \phi_n(a\tau_n + \tau_{ns}) + \phi_s\tau_s. \end{aligned} \tag{2.26}$$

Transport equation for nutrients.

$$\begin{aligned} &\frac{\partial}{\partial t}(\phi_s c) + \nabla\cdot(c\mathbf{v}\phi_s - D_s\phi_s\nabla c) = -g_c, \qquad \text{(CA-model)}\\ &\frac{\partial}{\partial t}(\phi_s c) + \nabla\cdot(c\mathbf{v}_s\phi_s - D_s\phi_s\nabla c) = -g_c. \qquad \text{(CN-model)} \end{aligned} \tag{2.27}$$

Transport equation for the polymer network volume fraction.

$$\begin{aligned} &\frac{\partial\phi_n}{\partial t} + \nabla\cdot(\phi_n\mathbf{v}) = \nabla\cdot\left[\lambda_{ch}\nabla\frac{\delta f}{\delta\phi_n}\right] + g_n, \qquad \text{(CH-model)}\\ &\frac{\partial\phi_n}{\partial t} + \nabla\cdot(\phi_n\mathbf{v}) = \nabla\cdot\left[\lambda\phi_n\nabla\frac{\delta f}{\delta\phi_n}\right] + g_n. \qquad \text{(MCH-model)} \end{aligned} \tag{2.28}$$

Constitutive equations.

$$\begin{gathered} \tau_n = 2\eta_n\mathbf{D},\ \tau_{ns} = 0,\ \tau_s = 2\eta_s\mathbf{D},\ a = 1, \qquad \text{(VA-model)}\\ \tau_n = 2\eta_n\mathbf{D}_n,\ \tau_{ns} = 0,\ \tau_s = 2\eta_s\mathbf{D}_s,\ a = 1, \qquad \text{(VN-model)}\\ \frac{d\tau_n}{dt} - \mathbf{W}\cdot\tau_n + \tau_n\cdot\mathbf{W} - a[\mathbf{D}\cdot\tau_n + \tau_n\cdot\mathbf{D}] + \frac{\tau_n}{\lambda_1} = \frac{2\eta_p}{\lambda_1}\mathbf{D},\\ \tau_{ns} = 2\eta_n\mathbf{D},\ \tau_s = 2\eta_s\mathbf{D}, \qquad \text{(JSA-model)}\\ \frac{\partial\tau_n}{\partial t} + \nabla\cdot(\mathbf{v}_n\tau_n) - \mathbf{W}_n\cdot\tau_n + \tau_n\cdot\mathbf{W}_n - a[\mathbf{D}_n\cdot\tau_n + \tau_n\cdot\mathbf{D}_n] + \frac{\tau_n}{\lambda_1} = \frac{2\eta_p}{\lambda_1}\mathbf{D}_n,\\ \tau_{ns} = 2\eta_n\mathbf{D}_n,\ \tau_s = 2\eta_s\mathbf{D}_s. \qquad \text{(JSN-model)} \end{gathered} \tag{2.29}$$

The production rate for polymer network and the consumption rate for the nutrient follow those of the Cogan–Keener model defined in section 1. In the MCH model, the mobility parameter λ can also be assigned to $\lambda_0\phi_s$ in case the solvent volume fraction is low and varies drastically in space.

3. Nondimensionalization. We use a characteristic time-scale t_0 and length-scale h, whose values will be specified in specific applications, to nondimensionalize the variables

$$\tilde{t} = \frac{t}{t_0},\quad \tilde{\mathbf{x}} = \frac{\mathbf{x}}{h},\quad \tilde{\mathbf{v}} = \frac{\mathbf{v}t_0}{h},\quad \tilde{p} = \frac{pt_0^2}{\rho_0 h^2},\quad \tilde{\tau}_n = \frac{\tau_n t_0^2}{\rho_0 h^2},\quad \tilde{c} = \frac{c}{c_0}, \tag{3.1}$$

where c_0 is a characteristic substrate concentration. The following dimensionless quantities arise:

(3.2)
$$\Lambda = \frac{\lambda \rho_0}{t_0},\ \Gamma_1 = \frac{\gamma_1 k T t_0^2}{\rho_0 h^4},\ \Gamma_2 = \frac{\gamma_2 k T t_0^2}{\rho_0 h^2},\ Re_s = \frac{\rho_0 h^2}{\eta_s t_0},\ Re_n = \frac{\rho_0 h^2}{\eta_n t_0},\ Re_p = \frac{\rho_0 h^2}{\eta_p t_0},$$
$$\tilde{D}_s = \frac{D_s t_0}{h^2},\ \Lambda_1 = \frac{\lambda_1}{t_0},\ \tilde{\rho} = \phi_s \frac{\rho_s}{\rho_0} + \phi_n \frac{\rho_n}{\rho_0},\ \tilde{A} = A t_0,\ \tilde{\mu} = \mu t_0,\ \tilde{K}_c = \frac{K_c}{c_0},$$

where ρ_0 is an averaged density; $Re_{s,n,p}$ are the Reynolds numbers for the solvent, bacteria in the effective EPS polymer network, and EPS polymer network; Λ_1 is the Deborah number for the polymer network; Λ, $\Gamma_{1,2}$, $\tilde{D}_s$, $\tilde{A}$, $\tilde{\mu}$, $\tilde{K}_c$ are the dimensionless parameters of their dimensional counterparts. For simplicity, we drop the $\tilde{\bullet}$ on the dimensionless variables, and the parameters. The system of governing equations in these dimensionless variables are given, for example in the CH+CA+JSA model, by

$$\begin{aligned}
&\nabla \cdot \mathbf{v} = 0, \\
&\rho \frac{d\mathbf{v}}{dt} = \nabla \cdot (\phi_n (a\tau_n + \tau_{ns}) + \phi_s \tau_s) - [\nabla p + \Gamma_1 \nabla \cdot (\nabla \phi_n \nabla \phi_n)], \\
&\frac{\partial}{\partial t}(\phi_s c) + \nabla \cdot (c\mathbf{v}\phi_s - D_s \phi_s \nabla c) = -g_c, \quad \text{(CA)} \\
&\frac{\partial \phi_n}{\partial t} + \nabla \cdot (\phi_n \mathbf{v}) = \nabla \cdot \left[\Lambda \nabla \frac{\delta f}{\delta \phi_n}\right] + gn, \quad \text{(CH)} \\
&\frac{d\tau_n}{dt} - \mathbf{W} \cdot \tau_n + \tau_n \cdot \mathbf{W} - a[\mathbf{D} \cdot \tau_n + \tau_n \cdot \mathbf{D}] + \frac{\tau_n}{\Lambda_1} = \frac{2}{\Lambda_1 Re_p} \mathbf{D}, \\
&\tau_{ns} = \frac{2}{Re_n} \mathbf{D}, \quad \tau_s = \frac{2}{Re_s} \mathbf{D}, \quad g_c = A\phi_n c, \quad gn = \epsilon \mu \phi_n \frac{c}{K_c + c}.
\end{aligned} \tag{3.3}$$

The dimensionless mixing free energy density is now given by

$$f = \frac{\Gamma_1}{2} \|\nabla \phi_n\|^2 + \Gamma_2 \left[\frac{\phi_n}{N} \ln \phi_n + (1 - \phi_n) \ln(1 - \phi_n) + \chi \phi_n (1 - \phi_n)\right]. \tag{3.4}$$

The other dimensionless equations can be obtained analogously. To save space, we will not enumerate them here.

4. Steady states in one dimension and their linear stability. In this section we examine the solution of the governing system of equations that depends on one spatial variable $y \in I = [0, 1]$, where the characteristic length-scale h is chosen as the width of the stripe which the fluid mixture occupies. The boundary conditions for the governing system of equations are

$$\mathbf{v}|_{\partial I} = \mathbf{v}^0, \quad \left[\phi_n \mathbf{n} \cdot \nabla \frac{\delta f}{\delta \phi_n}\right]_{\partial I} = 0, \quad [\mathbf{n} \cdot \nabla \phi_n]_{\partial I} = 0, \quad [\phi_s \mathbf{n} \cdot \nabla c]_{\partial I} = 0, \tag{4.1}$$

where $\mathbf{n}$ is the unit external normal at the boundary of the domain I and ∂I denotes the boundary of the domain. These boundary conditions consist of the no-slip boundary condition on the solid boundary for the average velocity, the no-penetration boundary condition for the excessive polymer network velocity, and a no-flux boundary condition for the polymer network volume fraction and for the nutrient concentration, respectively.

4.1. Viscous limit. We first discuss the solution given by the viscous model (CH+VA). Let $\eta_m = \frac{1-\phi_n}{Re_s} + \frac{\phi_n}{Re_n}$, where $1/\eta_m$ is the effective Reynolds number and

$$\hat{f}(\phi_n) = \frac{\phi_n}{N} \ln \phi_n + (1-\phi_n)\ln(1-\phi_n) + \chi\phi_n(1-\phi_n). \tag{4.2}$$

$\hat{f}$ is the bulk Flory–Huggins mixing free energy density. Considering the boundary condition at the wall, we set $v_y^0 = 0$.

The constant steady state solution for all models is given by

$$\begin{array}{llll} \mathbf{v} = \mathbf{0}, & p = p_0, & \phi_n = \phi_0, & c = 0, \text{ or} \\ \mathbf{v} = \mathbf{0}, & p = p_0, & \phi_n = 0, & c = c_0, \end{array} \tag{4.3}$$

where p_0 is an arbitrary constant, c_0 is an arbitrary positive constant, and $0 \leq \phi_0 < 1$ is a constant. In addition to the constant solutions, there can exist a nonconstant steady state at $c = 0$ for ϕ_n governed by

$$\Gamma_1 \phi_n'' - \Gamma_2 \frac{\partial \hat{f}}{\partial \phi_n} = \Gamma_1 C_0. \tag{4.4}$$

A closed form of the solution is not available for this equation. However, (4.4) can be integrated to yield

$$\phi_n' = \pm\sqrt{2C_0\phi_n + \frac{2\Gamma_2}{\Gamma_1}\hat{f}(\phi_n) + 2C_1}, \tag{4.5}$$

where C_0 and C_1 are integrating constants. A qualitative phase space analysis on an analogous system is given in [20]. Here we focus on the nonconstant steady state satisfying the Neumann boundary condition.

Using the boundary condition $\phi_n'(1) = \phi_n'(0) = 0$, we can determine C_0 and C_1:

$$\begin{aligned} 2C_0\phi_n(1) + \frac{2\Gamma_2}{\Gamma_1}\hat{f}(\phi_n(1)) + 2C_1 = 0, \\ 2C_0\phi_n(0) + \frac{2\Gamma_2}{\Gamma_1}\hat{f}(\phi_n(0)) + 2C_1 = 0. \end{aligned} \tag{4.6}$$

If $\phi_n(0) \neq \phi_n(1)$,

$$C_0 = \frac{\Gamma_2}{\Gamma_1}\,\frac{\hat{f}(\phi_n(0)) - \hat{f}(\phi_n(1))}{\phi_n(1) - \phi_n(0)}, \qquad C_1 = -\frac{\Gamma_2}{\Gamma_1}\,\frac{\phi_n(1)\hat{f}(\phi_n(0)) - \phi_n(0)\hat{f}(\phi_n(1))}{\phi_n(1) - \phi_n(0)}. \tag{4.7}$$

If we denote

$$g(\phi) = -\frac{\Gamma_1}{\Gamma_2}[C_0\phi_n + C_1], \tag{4.8}$$

$g(\phi)$ is the secant-line interpolating the points $(\phi_n(0), \hat{f}(\phi_n(0)))$ and $(\phi_n(1), \hat{f}(\phi_n(1)))$. In order to have a smooth real solution, $\hat{f} - g > 0$; i.e., $\hat{f}$ is concave down between $\phi_n(0)$ and $\phi_n(1)$. The concavity region of $\hat{f}$ is depicted in Figure 1 in phase space (ϕ, χ) at $N = 1000$. In the concave down region, a smooth solution can exist depending on the magnitude of $\frac{2\Gamma_2}{\Gamma_1}$.

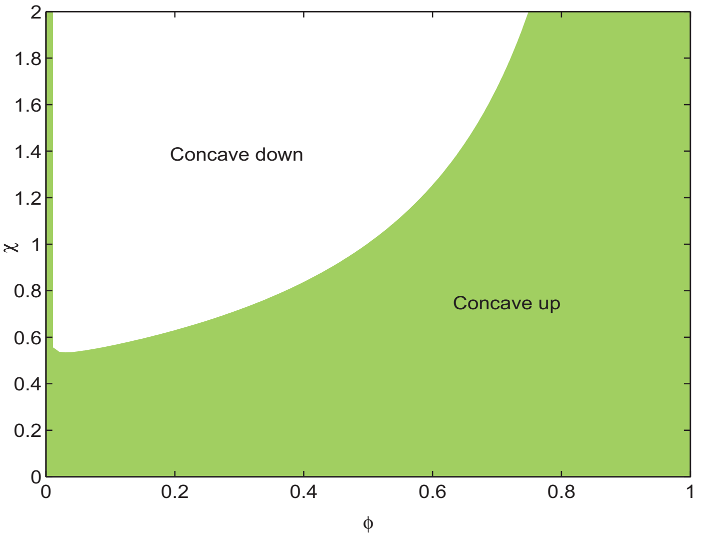


FIG. 1. *The regions of concavity in phase space* (ϕ, χ) *at* $N = 1000$*.*

From (4.5), we can see that the steady state solution is either monotonically increasing or decreasing if it exists. Integrating (4.5), we arrive at

$$\pm \int_{\phi_n(0)}^{\phi_n(y)} \frac{d\phi}{\sqrt{\hat{f}(\phi) - g(\phi)}} = y\sqrt{\frac{2\Gamma_2}{\Gamma_1}}, \tag{4.9}$$

where the solution of the boundary value problem is constrained by

$$\pm \int_{\phi_n(0)}^{\phi_n(1)} \frac{d\phi}{\sqrt{\hat{f}(\phi) - g(\phi)}} = \sqrt{\frac{2\Gamma_2}{\Gamma_1}}. \tag{4.10}$$

Notice that $\frac{2\Gamma_2}{\Gamma_1} = \frac{2h^2\gamma_2}{\gamma_1}$. Unless this dimensionless quantity is small, there could not be a solution to the integral equation. When the right-hand side is small, the chance to have a smooth solution increases considerably.

If $\phi_n(0) = \phi_n(1)$, we can determine C_0 only in terms of C_1:

$$C_1 = -C_0\phi_n(1) - \frac{\Gamma_2}{\Gamma_1}\hat{f}(\phi_n(1)). \tag{4.11}$$

The governing equation is given by

$$\phi_n' = \pm\sqrt{\frac{2\Gamma_2}{\Gamma_1}\left[\hat{f}(\phi_n) - \left(-\frac{C_0\Gamma_1}{\Gamma_2}(\phi_n - \phi_n(0)) + \hat{f}(\phi_n(0))\right)\right]}. \tag{4.12}$$

The constant solution $\phi_n = \phi_n(0)$ is a solution. When $\hat{f}$ is concave down, there could be a nonconstant steady state given below, provided that $\frac{2h^2\gamma_2}{\gamma_1}$ is small:

$$\begin{aligned}(4.13)\quad &\int_{\phi_n(0)}^{\phi_n(y)} \frac{d\phi}{\sqrt{\left[\hat{f}(\phi_n) - \left(-\frac{C_0\Gamma_1}{\Gamma_2}(\phi_n - \phi_n(0)) + \hat{f}(\phi_n(0))\right)\right]}} = y\sqrt{\frac{2\Gamma_2}{\Gamma_1}}, \quad 0 \le y \le \frac{1}{2},\\ &\int_{\phi_n(1/2)}^{\phi_n(y)} \frac{d\phi}{\sqrt{\left[\hat{f}(\phi_n) - \left(-\frac{C_0\Gamma_1}{\Gamma_2}(\phi_n - \phi_n(0)) + \hat{f}(\phi_n(0))\right)\right]}} = -y\sqrt{\frac{2\Gamma_2}{\Gamma_1}}, \quad \frac{1}{2} < y \le 1,\end{aligned}$$

or

$$\begin{aligned}(4.14)\quad &\int_{\phi_n(0)}^{\phi_n(y)} \frac{d\phi}{\sqrt{\left[\hat{f}(\phi_n) - \left(-\frac{C_0\Gamma_1}{\Gamma_2}(\phi_n - \phi_n(0)) + \hat{f}(\phi_n(0))\right)\right]}} = -y\sqrt{\frac{2\Gamma_2}{\Gamma_1}}, \quad 0 \le y \le \frac{1}{2},\\ &\int_{\phi_n(1/2)}^{\phi_n(y)} \frac{d\phi}{\sqrt{\left[\hat{f}(\phi_n) - \left(-\frac{C_0\Gamma_1}{\Gamma_2}(\phi_n - \phi_n(0)) + \hat{f}(\phi_n(0))\right)\right]}} = y\sqrt{\frac{2\Gamma_2}{\Gamma_1}}, \quad \frac{1}{2} < y \le 1.\end{aligned}$$

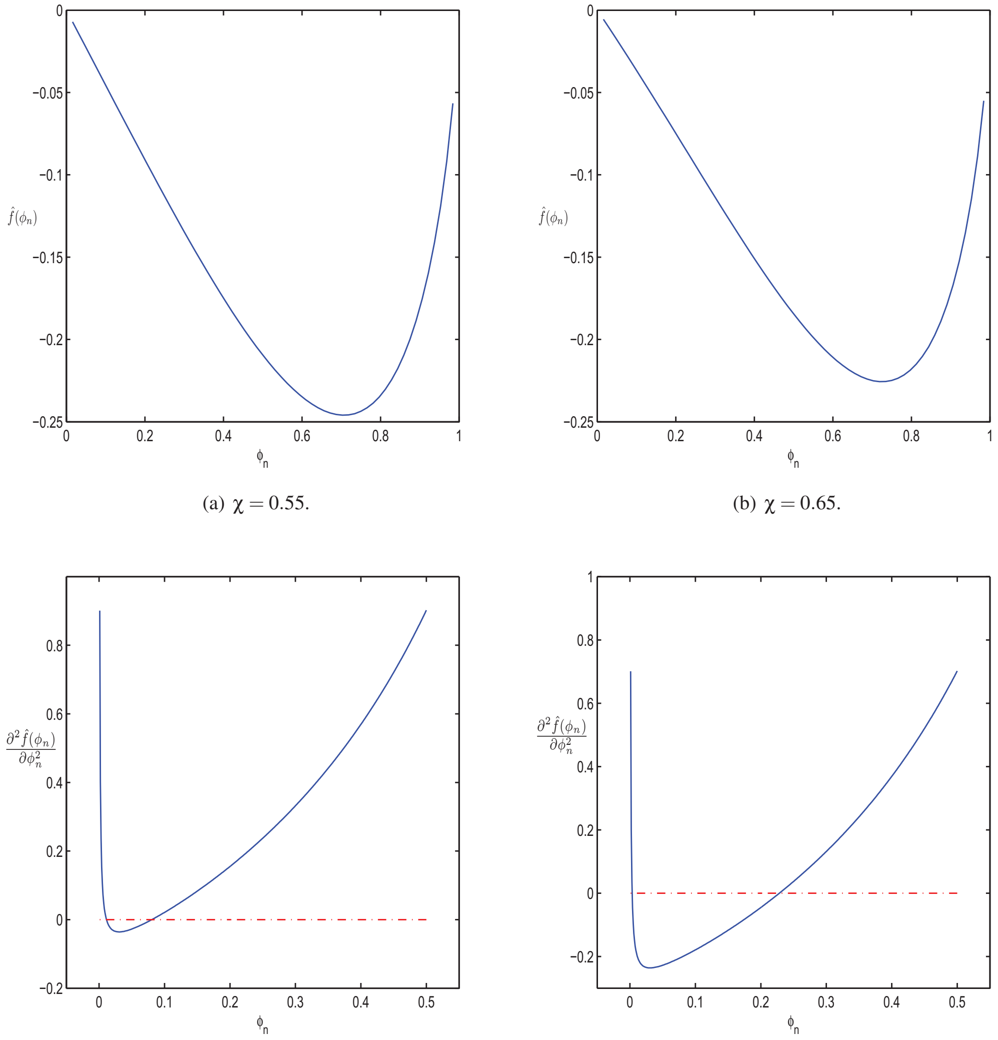
This solution is spatially periodic with period 1.

Next, we examine the linearized stability of the constant states. Let $\rho^0 = \rho(\phi_0)$ be the averaged density at the steady state. The eigenfunction for the velocity components is $\sin(\beta y)$ and for c and ϕ_n is $\cos(\beta y)$, respectively, where $\beta = m\pi$, $m = 1, \ldots, \infty$. The growth rates of the linearized system are given by

$$\begin{aligned}\alpha_{1,2} &= -\frac{1}{\rho^0}\left(\frac{1-\phi_0}{Re_s} + \frac{\phi_0}{Re_n}\right)\beta^2,\\ \alpha_3 &= \Lambda\left(-\Gamma_2\frac{\partial^2 \hat{f}}{\partial \phi^2}(\phi_0)\beta^2 - \Gamma_1\beta^4\right),\\ \alpha_4 &= -D_s\beta^2 - A\phi_0,\end{aligned} \tag{4.15}$$

where $\alpha_{1,2}$ are the growth rates obtained from the linearized momentum equations, α_3 is the growth rate corresponding to the linearized transport equation for ϕ_n, and α_4 is the growth rate for the nutrient concentration. If $\frac{\partial^2 \hat{f}}{\partial \phi^2}(\phi_0) \ge 0$, i.e., the bulk mixing energy density curve is concave up, all the growth rates are nonpositive; in fact, they are decay rates. Otherwise, in the portion where the mixing energy density is concave down, α_3 is positive for small values of β and negative for large values of β, i.e., the steady state suffers the long wave instability. We note that $\frac{\partial^2 \hat{f}}{\partial \phi_n^2} = \frac{1}{N\phi_n} + \frac{1}{1-\phi_n} - 2\chi$, and thus $\frac{\partial^2 \hat{f}}{\partial \phi_n^2} = 0$ has two solutions $\phi_n^{\pm}$. If $\phi_n^{\pm}$ are real, $\frac{\partial^2 \hat{f}}{\partial \phi_n^2} < 0$ and $\hat{f}$ is concave down for $\phi_n^- < \phi_n < \phi_n^+$. The instability occurs in the concave down region. Figure 2 depicts $\hat{f}$ and $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}$ at $N = 10^3$ and two different values of χ. In (c) and (d), the intersections of the dashed line with the curve give values of $\phi_n^{\pm}$. It can be seen for larger values of χ that the range of ϕ_n where $\frac{\partial^2 \hat{f}}{\partial \phi^2 n} < 0$ becomes wider.

For the second family of constant steady states (4.3.2) the eigenfunctions for the velocity, the nutrient substrate concentration, and the polymer network volume fraction are identical to the previous case given by either $\sin(\beta y)$ or $\cos(\beta y)$. The growth rates of the linearized system are given by



(a) $\chi = 0.55$.

(b) $\chi = 0.65$.

(c) $\chi = 0.55$.

(d) $\chi = 0.65$.

FIG. 2. *The normalized bulk mixing energy density* $\hat{f}(\phi_n)$ *and its second derivatives* $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}$ *at* $\chi = 0.55, 0.65$. *At* $\chi = 0.55$, *the concave down region does not include* $\phi_n = 0.19$, *whereas it does at* $\chi = 0.65$.

(4.16)
$$\alpha_{1,2} = -\frac{1}{\rho^0 Re_s}\beta^2, \quad \alpha_3 = \Lambda\left(-\Gamma_2 \frac{\partial^2 \hat{f}}{\partial \phi^2}(0)\beta^2 - \Gamma_1 \beta^4\right) + \frac{\epsilon \mu c_0}{K + c_0}, \quad \alpha_4 = -D_s \beta^2.$$

We note that $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}(0)$ is not defined in the original definition of the Flory–Huggins mixing free energy density. However, if we modify the $\phi_n \ln \phi_n$ term in the mixing energy density f by $(\phi_n + \delta\phi)\ln(\phi_n + \delta\phi)$, where $0 < \delta\phi \ll 1$, then we have

$$\frac{\partial^2 \hat{f}}{\partial \phi_n^2} = \frac{1}{N(\phi_n + \delta\phi)} + \frac{1}{1 - \phi_n} - 2\chi,$$

and $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}(0) = \frac{1}{N\delta\phi} + 1 - 2\chi$. Here, $\Delta\phi$ is a small positive number. If $\delta\phi \leq \frac{1}{N}$ and $0 \leq \chi \leq 1$, then $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}(0) \geq 0$ and the only positive growth rate comes from the polymer network production term at small β. For practical purposes, we use $\delta\phi = 10^{-6}$ throughout this paper.

We remark that the linearized stability analysis applies to the equations in an infinite domain and higher space dimensions as well. In this case, $\beta = \mathbf{k} \cdot \mathbf{l}$, where $\mathbf{k}$ is the wave number, $\mathbf{l}$ is a fixed direction in the multidimensional space, and the eigenfunctions are the Fourier (normal) modes. The analysis also applies to the three-dimensional cubic domain with homogeneous or periodic boundary conditions.

Figure 3 depicts the growth rates for the two families of constant steady states with dimensionless parameters $\Lambda = 10^{-9}$, $\Gamma_1 = 41.8337$, $\Gamma_2 = 418337$, $N = 10^3$, $\epsilon = 1$ and two selected values of χ at $\phi_n = 0.19$. For the first family of constant steady states, when $\chi = 0.55$, Figure 2(c) shows $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}(0.19) > 0$, and thus the growth rate $\alpha_3 < 0$ for all $\beta > 0$; when $\chi = 0.65$, Figure 2(d) shows $\frac{\partial^2 \hat{f}}{\partial \phi_n^2}(0.19) < 0$, and thus $\alpha_3 > 0$ for β between 0 and approximately 24. For the second family of constant steady states, a long wave instability persists to the infinitely long wave limit at any χ. Numerical results confirming the long wave instability in nonlinear regimes are presented in section 6.

For the MCH model, the growth rate α_3 is simply modified by a factor of ϕ_0 for the first family of constant steady states

$$\alpha_3 = \phi_0 \Lambda \left(-\Gamma_2 \frac{\partial^2 \hat{f}}{\partial \phi^2}(\phi_0)\beta^2 - \Gamma_1 \beta^4 \right), \tag{4.17}$$

whereas it is given by

$$\alpha_3 = \frac{\epsilon \mu c_0}{K_c + c_0} \tag{4.18}$$

for the second family of constant steady states, which equals the infinitely long wave limit of the growth rate α_3 in (4.16).

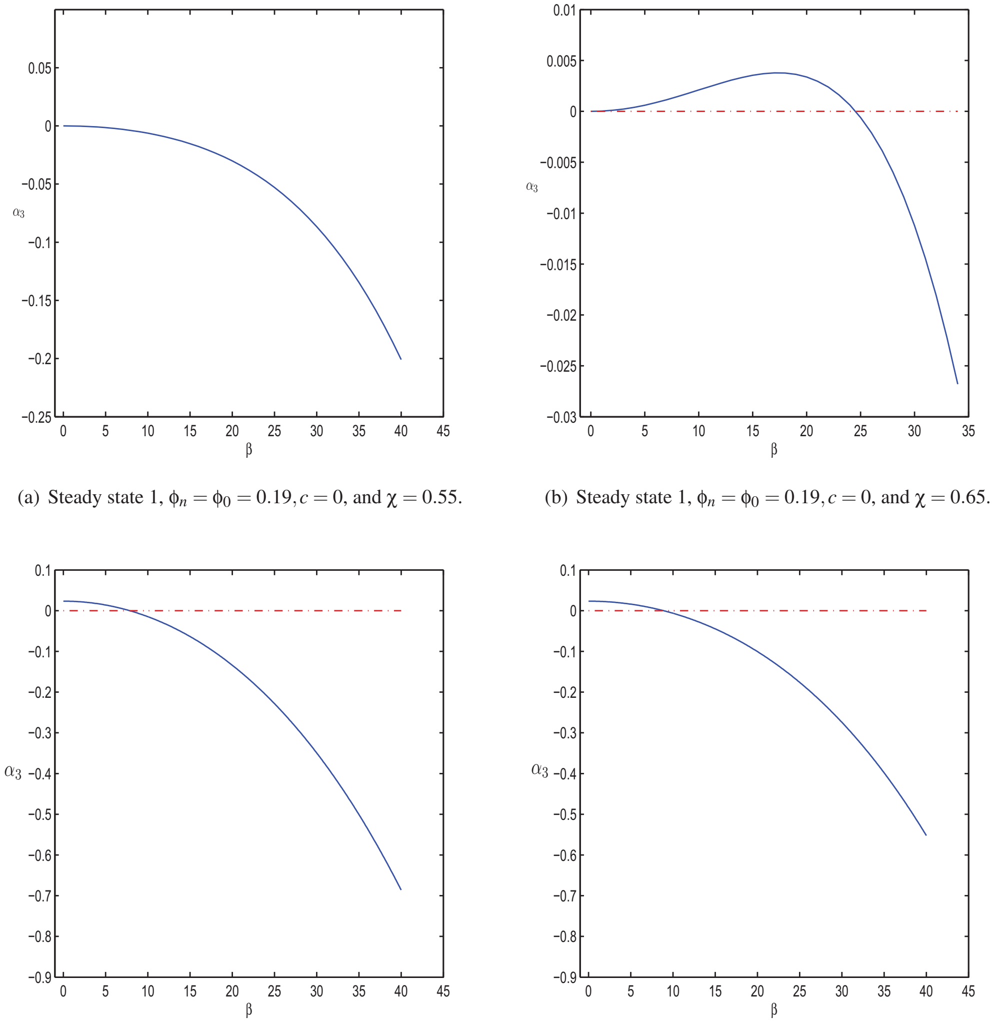
We next examine the steady states and their stability in the viscoelastic models.

4.2. Viscoelastic model. The viscoelastic model adds a set of constitutive equations for the elastic stress to the governing system of equations and couples the elastic stress to the momentum transport equation. For brevity, we use τ in place of τ_n from here on for the polymer elastic stress tensor.

Notice that the steady state of the elastic stress tensor is zero; the constitutive equation for the polymer network stress is independent of the volume fraction ϕ_n and concentration c; given the zero boundary conditions on $\mathbf{v}$, it is not necessary to impose any boundary conditions on the polymer elastic stress components. We obtain that four modes in the linearized constitutive equation are independent, and their growth rates are given by

$$\alpha_{5,7,8,10} = -\frac{1}{\Lambda_1}, \tag{4.19}$$

where the indices track the four decoupled modes of the elastic stress tensor. The other two modes $\alpha_{6,9}$ are coupled to the momentum equation. For the first family of steady states $\phi = \phi_0$, $c = 0$, the coupled growth rates are calculated as



(a) Steady state 1, $\phi_n = \phi_0 = 0.19, c = 0$, and $\chi = 0.55$.

(b) Steady state 1, $\phi_n = \phi_0 = 0.19, c = 0$, and $\chi = 0.65$.

(c) Steady state 2, $\phi_n = 0, c = c_0$, and $\chi = 0.55$.

(d) Steady state 2, $\phi_n = 0, c = c_0$, and $\chi = 0.65$.

FIG. 3. *Growth rate of the linearized CH model. The values of the dimensionless parameters are* $\Lambda = 10^{-9}$, $\Gamma_1 = 41.8337$, $\Gamma_2 = 418337$, $N = 10^3$, $\delta\phi = 10^{-3}$, $\epsilon = 1$, $c_0 = 0.1$, $\mu = 0.14$, $K_c = 0.5$. *For the first family of steady states, the long wave growth is due to the polymer-solvent mixing kinetics shown in* (b). *Panel* (a) *depicts a negative growth rate. In contrast, for the second family of steady states, the long wave growth rate depends only on the polymer production shown in* (c) *and* (d).

(4.20)

$$\alpha_{1,2,6,9} = \frac{1}{2\rho_0}\left[-\left(\frac{\rho_0}{\Lambda_1} + \left(\frac{1-\phi_0}{Re_s} + \frac{\phi_0}{Re_n}\right)\beta^2\right) \pm\sqrt{\left(\frac{\rho_0}{\Lambda_1} + \frac{1-\phi_0}{Re_s}\beta^2\right)^2 - 4\rho_0\left(\left(\frac{1-\phi_0}{\Lambda_1 Re_s} + \frac{\phi_0}{\Lambda_1 Re_n}\right) + \frac{2a\phi_0}{\Lambda_1 Re_p}\beta^2\right)}\,\right].$$

The rates all have negative real parts. The corresponding eigenfunction for the velocity components is $\sin\beta y$, and that for the stress components is $\cos\beta y$. The growth rates

$\alpha_{3,4}$ and eigenfunctions for ϕ_n and c are identical to those in the viscous limit.

For the second family of steady states $\phi = 0$, $c = c_0$. The linearized momentum and constitutive equations decouple. So, the growth rates in the viscous limit $\alpha_{1,2}$ remain in addition to the decay rates from the constitutive equations,

$$\alpha_{5,6,7,8,9,10} = -\frac{1}{\Lambda_1}. \tag{4.21}$$

In the gel model ($\Lambda_1 \to \infty$), the growth rates for the first family of steady states are given by

$$\begin{aligned} \alpha_{1,2} &= -\left(\frac{1-\phi_0}{Re_s} + \frac{\phi_0}{Re_n}\right)\beta^2, \\ \alpha_{5,6,7,8,9,10} &= 0. \end{aligned} \tag{4.22}$$

The results for the JSN model are qualitatively the same and are omitted here. The analysis shows that the viscoelasticity at the linear regime does not have any negative effects on the stability. We next study the nonlinear dynamics of the biofilm flows in one space dimension. But first we present the numerical method that we use to compute the nonlinear transient solutions.

5. Numerical scheme for the one-dimensional biofilm models. In this section we investigate the growth of the biofilm in one dimension: $y \in I = [0, 1]$ governed by the momentum, Cahn–Hilliard or modified Cahn–Hilliard equations, the nutrient transport equation, and the stress constitutive equation JSA or JSN with the continuous supply of nutrient substrates through the top boundary. We adopt the boundary conditions given in (4.1) except that the nutrient boundary conditions are replaced by

$$[D\phi_s \nabla_y c] \cdot \mathbf{n}|_{y=0} = 0, \qquad c|_{y=1} = c^\star, \tag{5.1}$$

where $\mathbf{n}$ is the unit outward normal of domain I. The boundary condition of c at $y = 1$ is the Dirichlet one, $c|_{y=1}\ c^\star$, indicating that the substrate is fed at the top boundary to maintain a constant level of $c = c^\star$. The boundary condition for the velocity is chosen to be $\mathbf{v}_0|_{y=0} = (0, 0, 0)^T$, $\mathbf{v}_0|_{y=1} = (10^{-3}, 0, 0)^T$. We note that the vanishing boundary condition for v_y along with the continuity condition warrants a vanishing velocity component in the y direction. Thus the transport of the polymer network is entirely due to the excessive flux.

The numerical scheme used to study the dynamics of biofilm growth is a finite difference scheme. We use uniform spatial and time step sizes, denoted by Δy and Δt, respectively. For given solutions at time step $n-1$ and n the polymer volume fraction at time step $n+1$, ϕ_n^{n+1} governed by the Cahn–Hilliard equation is calculated by

$$\begin{aligned} &\frac{\phi_n^{n+1} - \phi_n^n}{\Delta t} + \theta\Lambda\nabla_y^2[\Gamma_1\nabla_y^2\phi_n^{n+1} + 2\Gamma_2\chi\phi_n^{n+1}] \\ &\quad = g_n(\bar{\phi}_n^{n+\theta}, \bar{c}^{n+\theta}) - (1-\theta)\Lambda\nabla_y^2[\Gamma_1\nabla_y^2\phi_n^n + 2\Gamma_2\chi\phi_n^n] \\ &\qquad - \Lambda\nabla_y^2\Gamma_2\left(-\frac{1}{N}\ln\bar{\phi}_n^{n+\theta} + \ln(1-\bar{\phi}_n^{n+\theta})\right). \end{aligned} \tag{5.2}$$

After this, the volume fraction of the solvent at time step $n+1$ is obtained by $\phi_s^{n+1} = 1 - \phi_n^{n+1}$, and the nutrient substrate concentration at time step $n+1$, c^{n+1} is calculated by

$$
\begin{aligned}
(5.3)\qquad &\frac{\phi_s^{n+1}c^{n+1}-\phi_s^n c^n}{\Delta t}-\theta\nabla_y\cdot(D_s\phi_s^{n+1}\nabla_y c^{n+1}-\mathbf{v}^{n+1}\phi_s^{n+1}c^{n+1})\\
&\quad=-g_c(\bar{\phi}_n^{n+\theta},\bar{c}^{n+\theta})+(1-\theta)\nabla_y\cdot(D_s\phi_s^n\nabla_y c^n-\mathbf{v}^n\phi_s^n c^n).
\end{aligned}
$$

The θ-method is used in time discretization of both equations, where $0\le\theta\le 1$, and the spatial discretization is done using central differences to ensure the second order accuracy in space and volume preservation for ϕ_n when there is no polymer production. Here, $\bar{\phi}_n^{n+\theta}=(1+\theta)\phi_n^n-\theta\phi_n^{n+1}$, $\bar{c}^{n+\theta}=(1+\theta)c^n-\theta c^{n+1}$ are the extrapolated values of ϕ_n and c at time step $n+\theta$, and the nonlinear functions g_n, g_c and the terms involving log-functions are evaluated at these extrapolated values. In our simulation throughout the paper, we use $\theta=1/2$, and thus the overall scheme is second order in time and space. The MCH equation is discretized similarly by

$$
\begin{aligned}
&\frac{\phi_n^{n+1}-\phi_n^n}{\Delta t}+\theta\Lambda\nabla\cdot[\bar{\phi}_n^{n+\theta}\nabla_y(\Gamma_1\nabla_y^2\phi_n^{n+1}+2\Gamma_2\chi\phi_n^{n+1})]\\
(5.4)\qquad &\quad=g_n(\bar{\phi}_n^{n+\theta},\bar{c}^{n+\theta})-(1-\theta)\Lambda\nabla\cdot[\phi_n^n\nabla_y(\Gamma_1\nabla_y^2\phi_n^n+2\Gamma_2\chi\phi_n^n)]\\
&\qquad-\Lambda\Gamma_2\nabla\cdot\left[\phi_n^n\nabla_y\left(-\frac{1}{N}\ln\bar{\phi}_n^{n+\theta}+\ln(1-\bar{\phi}_n^{n+\theta})\right)\right].
\end{aligned}
$$

Assuming that interval $I=[0,1]$ is divided into M uniform subintervals of size $\Delta y=1/M$ by $M+1$ nodes $y_0,y_1,\dots,y_M$, we denote the value of the numerical solution of (5.2) and (5.3) at $(n\Delta t, j\Delta y)$ by $\phi_{n,j}^n$, c_j^n, $j=0,\dots,M$. Since $\mathbf{v}\cdot\mathbf{n}|_{\partial I}=\mathbf{v}_0\cdot\mathbf{n}=0$, the discrete form of the boundary conditions (5.1) is given by

$$
\begin{aligned}
(5.5)\qquad &\phi_{n,1}^n=\phi_{n,-1}^n,\quad \phi_{n,2}^n=\phi_{n,-2}^n,\quad \phi_{n,M+1}^n=\phi_{n,M-1}^n,\quad \phi_{n,M+2}^n=\phi_{n,M-2}^n,\\
&c_1^n=c_{-1}^n,\quad c_M^n=c^\star.
\end{aligned}
$$

For the purpose of completeness, we also compute the nonzero velocity components v_x, v_z and the stress components $\tau_{xx},\tau_{xy},\dots,\tau_{zz}$, even though they are driven by ϕ_n and c. The time discretization of the equation for v_x is given by

$$
\begin{aligned}
&\rho^{n+1}\frac{v_x^{n+1}-v_x^n}{\Delta t}-\theta\frac{\partial}{\partial y}\left(\left(\frac{\phi_s^{n+1}}{Re_s}+\frac{\phi_n^{n+1}}{Re_p}\right)\frac{\partial v_x^{n+1}}{\partial y}\right)\\
(5.6)\qquad &\quad=(1-\theta)\frac{\partial}{\partial y}\left(\left(\frac{\phi_s^n}{Re_s}+\frac{\phi_n^n}{Re_p}\right)\frac{\partial v_x^n}{\partial y}\right)+\frac{\partial(a\phi_n^n\tau_{xy}^n)}{\partial y}.
\end{aligned}
$$

The spatial discretization is again central difference. The discrete equation for v_z is done similarly. Dirichlet boundary conditions are imposed for v_x and v_z; i.e., $v_{x,0}^n$, $v_{x,M}^n$, $v_{z,0}^n$, $v_{z,M}^n$ are given.

We note that all six components of the stress tensor satisfy a generic equation of the form

$$
(5.7)\qquad \frac{\partial\tau}{\partial t}+v_y\frac{\partial\tau}{\partial y}=F(\tau,\nabla\mathbf{v}).
$$

Here $F(\tau,\mathbf{v})$ has different forms for different components of the stress tensor, and it does not contain terms involving partial derivatives of τ. We also note that $\mathbf{v}$ can be either the polymer network velocity (JSN) (the sum of the average velocity and the excessive velocity) or the average velocity (JSA), depending on the model we choose. In the following, we adopt the constitutive model using the polymer network velocity. Since $v_y = 0$ at $y = 0, 1$, there are no boundary conditions for the elastic stress tensor τ; thus, τ actually satisfies an ODE: $\frac{\partial \tau}{\partial t} = F(\tau, \nabla \mathbf{v})$ at $y = 0, 1$. Then at the discrete level, we solve τ_0, τ_M by the following Runge–Kutta method:

$$\tau^{n+1} = \tau^n + \frac{\Delta t}{6}(K_1 + 2K_2 + K_3 + K_4), \tag{5.8}$$

where

$$K_1 = F(\tau^n, \nabla \mathbf{v}^n), \qquad K_2 = F\left(\tau^n + \frac{\Delta t}{2}K_1, \nabla\left(\frac{\mathbf{v}^n + \mathbf{v}^{n+1}}{2}\right)\right),$$

$$K_3 = F\left(\tau^n + \frac{\Delta t}{2}K_2, \nabla\left(\frac{\mathbf{v}^n + \mathbf{v}^{n+1}}{2}\right)\right), \qquad K_4 = F(\tau^n + \Delta t K_3, \nabla \mathbf{v}^{n+1}).$$

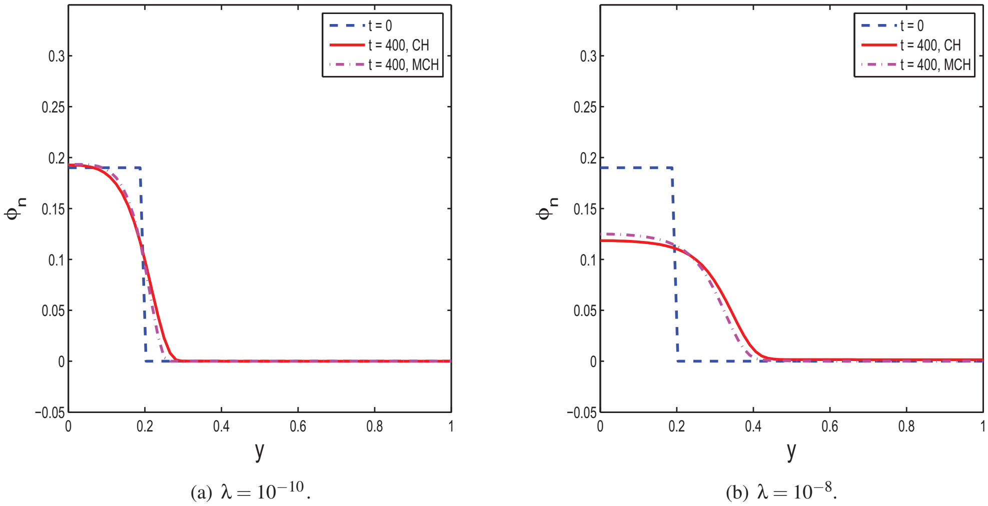
Away from the boundaries, we solve τ_j^n, $1 \leq j \leq M - 1$, by the following upwind scheme:

$$\begin{aligned} \frac{\tau_j^{n+1} - \tau_j^n}{\Delta t} = -\frac{1}{2\Delta y}\Big\{&[1 - \operatorname{sign}(v_{y,j+1/2}^n)]v_{y,j+1/2}^n(\tau_{j+1}^n - \tau_j^n) \\ &+ [1 + \operatorname{sign}(v_{y,j-1/2}^n)]v_{y,j-1/2}^n(\tau_j^n - \tau_{j-1}^n)\Big\} + F(\tau_j^n, \nabla \mathbf{v}^{\mathbf{n}}). \end{aligned} \tag{5.9}$$

6. Numerical results and dynamics of one-dimensional biofilms. We study the expansion and growth of one-dimensional biofilms that are homogeneous in the (x, z) plane and confined to the range $0 \leq y \leq 1$ using the numerical scheme developed in the previous section. Table 1 lists the values of the dimensional parameters used in our simulations [7]. In the phase field model, the mobility of the polymer network is assumed a material parameter, whose value can only be calibrated through material characterization in vitro or in vivo. In this numerical study, however, we

TABLE 1
Parameter values used in the simulation.

Symbol	Parameter	Value	Unit
T	Temperature	303	Kelvin
γ_1	Distortional energy	1×10^7	m kg s^{-2}
γ_2	Mixing free energy	1×10^{17}	m^{-1} kg s^{-2}
χ	Flory–Huggins parameter	0.55 or 0.65	dimensionless
N	Generalized polymerization parameter	1×10^3	dimensionless
μ	Max. production rate	1.4×10^{-4}	$\mathrm{kgm}^{-3}\mathrm{s}^{-1}$
K_c	Half saturation constant	5×10^{-4}	kgm^{-3}
A	Max. consumption rate	1	$\mathrm{kgm}^{-3}\mathrm{s}^{-1}$
D_s	Substrate diffusion coefficient	2.3×10^{-9}	$\mathrm{m}^2\mathrm{s}^{-1}$
η_n	Viscosity due to bacteria	4.3×10^2	$\mathrm{kgm}^{-1}\mathrm{s}^{-1}$
η_p	EPS polymer network viscosity	4.3	$\mathrm{kgm}^{-1}\mathrm{s}^{-1}$
η_s	Dynamic viscosity of solvent	1.002×10^{-3}	$\mathrm{kgm}^{-1}\mathrm{s}^{-1}$
ρ_n	Network density	1×10^3	kgm^{-3}
ρ_s	Solvent density	1×10^3	kgm^{-3}
c_0	Characteristic substrate concentration	1×10^{-3}	kgm^{-3}
h	Characteristic length-scale	1×10^{-3}	m
t_0	Characteristic time-scale	1×10^3	s
a	Slip parameter	0.92	dimensionless
M	Number of spacial subintervals	64	dimensionless



(a) $\lambda = 10^{-10}$.

(b) $\lambda = 10^{-8}$.

Fig. 4. *Evolution of the polymer volume fraction in one-dimensional biofilms by the CH and MCH models without the polymer production.* $\Delta t = \Delta y$ *in* (a), $\Delta t = 0.1\Delta y$ *in* (b). *The solution is plotted at* $t = 400$. *Clearly larger mobility transports the polymer network away from the high concentration regime, reducing the volume fraction of polymers there through conservation.*

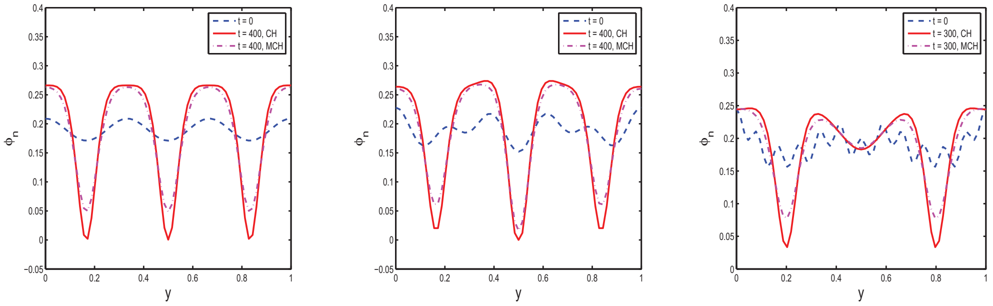
treat it as an operating parameter. Our first attempt is to characterize the effect of mobility on the dynamics of biofilms for both the CH and MCH models without the polymer production, i.e., $\epsilon = 0$, in which the volume fraction dynamics decouple from that of the nutrient substrate concentration. We then examine the variation of the mobility in the CH and MCH models when the polymer network production is present to select the appropriate model for our study of biofilm expansion and growth.

6.1. Biofilm dynamics with negligible EPS production. We begin with an initial profile of the polymer volume fraction distribution as a step function with a nonzero value at the bottom side of the domain and zero at the other side, e.g., $\phi_n(0, y) = 0.19$ for $0 \leq y \leq 0.2$, $\phi_n(0, y) = 0$ for $0.2 < y \leq 1$. This mimics the existence of a flat layer of biofilms in a gap of thickness 1 initially. Figure 4 depicts the evolution of the polymer volume fraction in the one-dimensional biofilm according to the CH model (5.2) for different values of mobility λ, in which the horizontal axis is y and the vertical axis is ϕ_n. In each plot, the step curve is the initial profile of ϕ_n at $t = 0$, and the solid smooth curve is ϕ_n at $t = 400$. Here, we choose the characteristic time-scale $t_0 = 1000$ seconds, so the dimensionless time $t = 400$ corresponds to about 4.6 days. λ as the mobility parameter controls the magnitude of the excessive flux for the polymer network due to polymer solvent mixing. We observe that when λ is small ($\lambda = 10^{-11} \sim 10^{-10}$), the effect of the excessive flux is small and the ϕ_n profile is only smoothed around the initial sharp interface (discontinuity) with a slightly accumulative expansion and growth, since the excessive polymer flux is not fast enough to transport the biomass out of the active mixing region. However, as λ increases to ($10^{-9} \sim 10^{-8}$), the biomass of the polymer network is transported rapidly to the nearby polymer-scarce region leading to sizable expansion of biofilms in the domain. We note that the no-flux boundary condition for ϕ_n at $y = 0$ and $y = 1$ leads to the total amount of conservation in ϕ_n, i.e., $\int_0^1 \phi_n(t, y)dy =$ const. Accompanying the sizable expansion of the biofilm, the volume fraction of the polymer network reduces in the nonzero ϕ_n (or biofilm) region at larger mobility due to this conservation property.

As a comparison, Figure 4 also plots the time evolution of the polymer volume fraction in one-dimensional biofilms according to the MCH model (5.4) for a comparable set of mobility values of λ rescaled by $\lambda \to \lambda/0.19$ to match the amount of polymeric fluxes in both models initially. The solutions of the MCH model are depicted in dot-dashed curves in the figure; they are qualitatively the same as those predicted using the CH model. However, there exists a subtle difference in the solution profile in ϕ_n in that the transport effect of the modified Cahn–Hilliard equation (5.4) outside the polymer-rich region is much weaker than that of the Cahn–Hilliard dynamics (5.2). This is due to the fact that the excessive polymeric flux in the MCH model is given by $-\lambda\phi_n\nabla\frac{\delta f}{\delta\phi_n}$ and vanishes when $\phi_n = 0$. On the other hand, the excessive flux in the CH model is given by $-\lambda\nabla\frac{\delta f}{\delta\phi_n}$ and may not be zero even if $\phi_n = 0$ due to the dissipative property of the CH equation and the numerical error. For example, in the case of $\lambda = 10^{-10}/0.19$, at $t = 400$, the value of ϕ_n at $y = 1$ is equal to 0 for the MCH model, and it is about 6×10^{-5} for the CH model at $\lambda = 10^{-10}$. This shows that the modified Cahn–Hilliard dynamics gives a much sharper excessive flux estimation in the solvent region than the Cahn–Hilliard dynamics does, and it also maintains a sharper interface between the biofilm and the solvent. This subtlety will be amplified in the following numerical studies when the polymer network production is accounted for.

Next we study the dynamics of the biofilm expansion without polymer network production ($\epsilon = 0$) in the neighborhood of constant steady states, considering an initial polymer volume fraction profile that is the perturbation from a constant steady state, e.g., $\phi_n(0, y) = 0.19 + 0.019\cos(2\pi k y)$, where k is the wave number. This transient simulation aims to investigate the nonlinear evolution of the constant steady states perturbed by either linearly stable or unstable modes. We choose two wave numbers: one falls into the linearly stable range ($k = 5$) and the other into the unstable range ($k = 3$) at $\chi = 0.65$. Our simulations demonstrate that the transient solutions corresponding to the linearly stable modes all converge to the homogenized steady state $\phi_0 = 0.19$, while the initial polymer volume fraction with the perturbation corresponding to the unstable mode evolves into a spatially inhomogeneous profile. Figure 5(a) depicts the polymer volume fraction profile corresponding to an unstable mode ($k = 3$) at $t = 400$. Since the difference in the stability between the CH and MCH models is seen only in the magnitude of the linearized growth rate, the results obtained from both models are qualitatively the same. The variation of volume fraction in the MCH model is smaller than that in the CH model though. Coarsening is observed in the transient simulation.

To further illustrate the nonlinear dynamics of the biofilm in the range of unstable wave numbers for $\chi = 0.65$, we investigate the evolution of the biofilm with initial polymer volume fraction of a perturbation of two different wave numbers: $\phi_0(y) = 0.19 + 0.019[\cos(2\pi \cdot 3 \cdot y) + \cos(2\pi \cdot 5 \cdot y)]$, where the perturbation of the constant steady state $\phi_0 = 0.19$ contains a growth mode $k = 3$ and a decay mode $k = 5$. Figure 5(b) depicts the numerical result at $t = 400$ for both the CH and MCH models, where the shorter wave mode ($k = 5$) decays and the longer one ($k = 3$) survives and grows, confirming the linear stability analysis. The nonlinear profile calculated from the MCH model is comparable to that from the CH model at the rescaled mobility parameter shown in Figure 5(b). Figure 5(c) portrays the evolution of the polymer volume fraction with initial condition given by a superposition of four different modes:



(a) one mode.

(b) two modes.

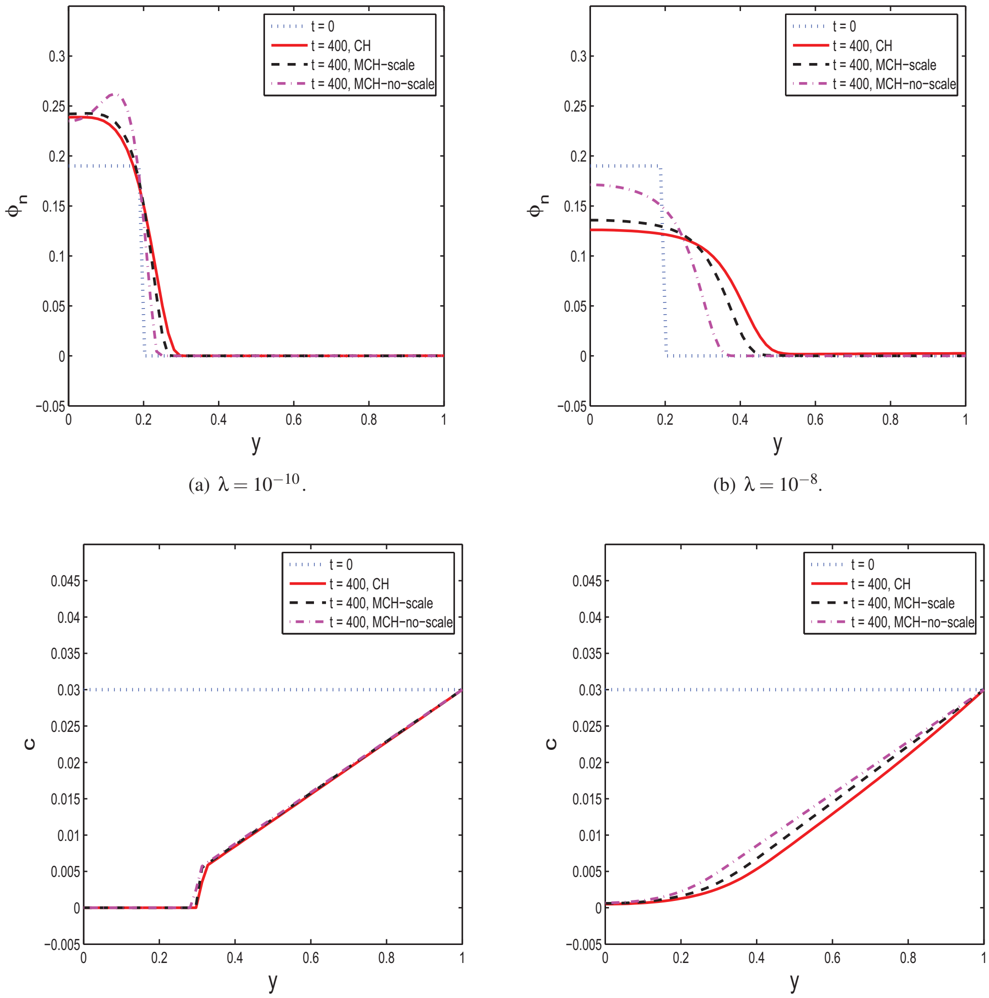
(c) four modes.

FIG. 5. *Evolution of the polymer volume fraction ϕ_n in one-dimensional biofilms by the CH and MCH models without polymer production at $\chi = 0.65$. The parameter $\lambda = 10^{-9}$.* (a) *The initial profile is given by $\phi_0(y) = 0.19 + 0.019\cos(2\pi k y)$, where $k = 3$. The polymer volume fraction tends to evolve (or coarsen) into islands with length-scale proportional to $1/k$.* (b) *The initial profile is given by $\phi_0(y) = 0.19 + 0.019[\cos(2\pi \cdot 3 \cdot y) + \cos(2\pi \cdot 5 \cdot y)]$.* (c) *The initial profile is given by $\phi_0(y) = 0.19 + 0.019[\xi_1 \cos(2\pi \cdot 2 \cdot y) + \xi_2 \cos(2\pi \cdot 3 \cdot y) + \xi_3 \cos(2\pi \cdot 5 \cdot y) + \xi_4 \cos(2\pi \cdot 12 \cdot y)]$, where ξ_i, $i = 1, \dots, 4$, are four randomly chosen constants.*

$\phi_0(y) = 0.19 + 0.019[\xi_1 \cos(2\pi \cdot 2 \cdot y) + \xi_2 \cos(2\pi \cdot 3 \cdot y) + \xi_3 \cos(2\pi \cdot 5 \cdot y) + \xi_4 \cos(2\pi \cdot 12 \cdot y)]$, where ξ_i, $1 \leq i \leq 4$, are random numbers chosen between 0 and 1. Here the perturbation contains two growth modes $k = 2, 3$ and two decay ones $k = 5, 12$. We observe that for both the CH and MCH models, the shorter waves ($k = 5, 12$) decay and the longer ones ($k = 2, 3$) grow. The profile of ϕ_n at $t = 300$ is a combination of the two "nonlinear modes" corresponding to $k = 2$ and $k = 3$, and the mode with $k = 3$ seems to be dominant, especially near the boundary. Note that $\beta = 2\pi k$, $k = 2$ and 3, correspond to $\beta = 12.57$ and 18.85, respectively. Figure 3(b) in section 4 indicates that the growth rate for $k = 3$ is bigger than that for $k = 2$, and thus our numerical results simply illustrate that the linear instability amplifies in the nonlinear regime.

6.2. Biofilm dynamics with EPS production in weak shear. Next, we turn to the growth case ($\epsilon = 1$) and study the expansion and growth of the biofilm with an initial profile of a step function in weak shear. The dimensionless shear speed at $y = 1$ is fixed at $v_x = 0.001$. The initial condition of the nutrient concentration is set at $c = c^\star = 0.03$ for $0 \leq y \leq 1$. Figure 6 depicts the results for the CH and MCH models. The step profile is ϕ_n at $t = 0$, and the smooth ones are ϕ_n at $t = 400$ obtained from both models. For the CH model, we observe that for small λ ($\lambda = 10^{-11} \sim 10^{-10}$), since the excessive flux is small, the polymer network mostly grows at the position where it is initially positive, and only a very small amount is transported to the right. It is also seen that the polymer network grows more rapidly around the interface between the biomass (mixture of polymer and solvent) and the pure solvent. This is because the nutrient to the left of the interface tends to be consumed in a short period of the film growth so as to cause the polymer network growth to cease after that, but the polymer around the interface can always access the nutrient due to the nutrient diffusion at the interface. Thus, the growth near the interface can be sustained. As λ increases (in $10^{-9} \sim 10^{-8}$), the polymer network expands into the solvent region, leading to a lower polymer volume fraction in the biofilm.

As a comparison, we repeat the same calculations using the MCH model at the



(a) $\lambda = 10^{-10}$.

(b) $\lambda = 10^{-8}$.

(c) $\lambda = 10^{-10}$.

(d) $\lambda = 10^{-8}$.

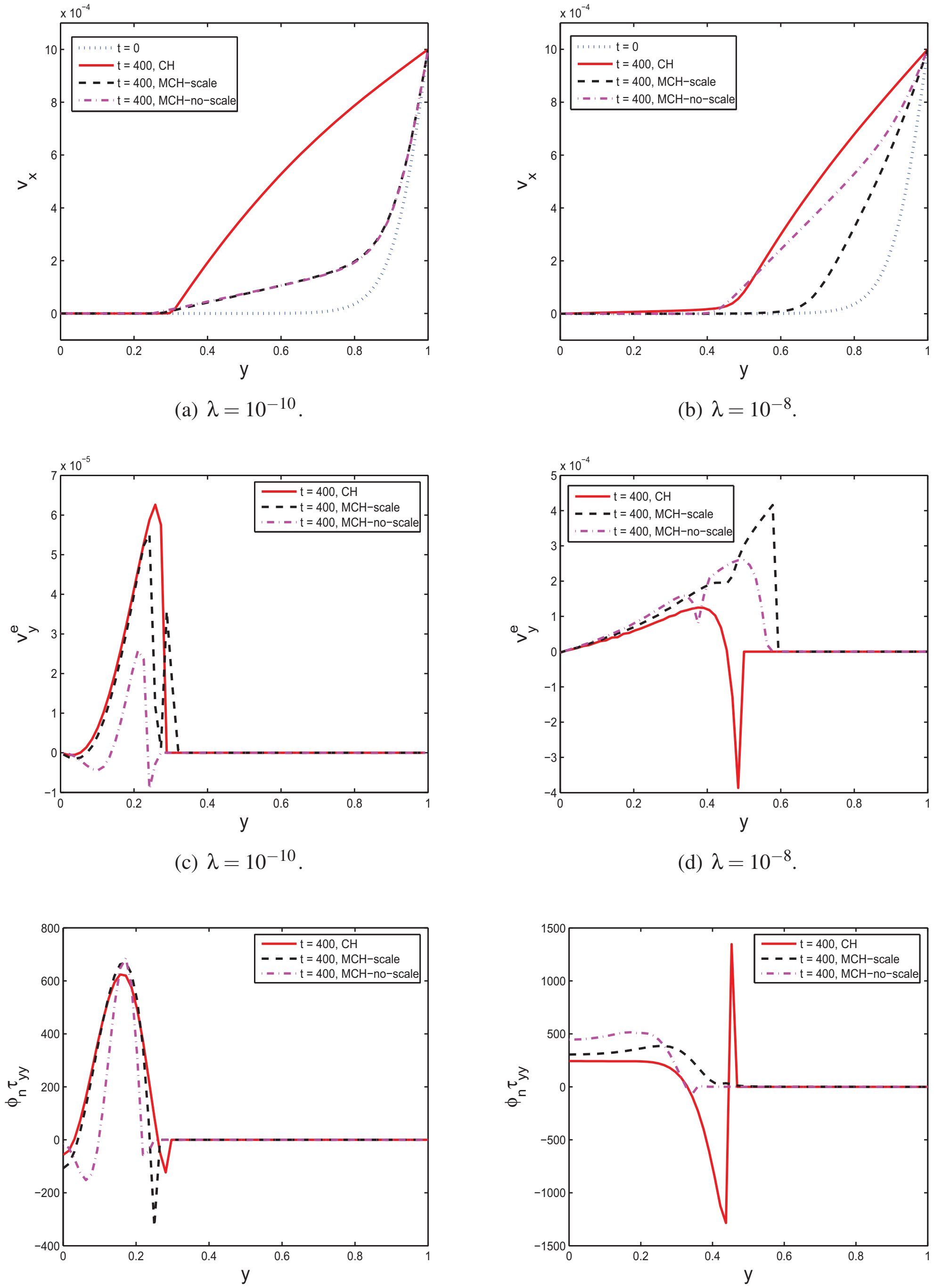
Fig. 6. *Growth of the polymer volume fraction in one-dimensional biofilms and the nutrient concentration profile. The parameter values are $\epsilon = 1$, $\mu = 0.14$, $K_c = 0.5$. The biofilm-solvent interface predicted by the MCH model always falls to the left of that predicted by the CH model. In addition, the MCH model gives a more realistic estimation of the volume fraction away from the biofilm in the solvent region and allows a slightly richer supply of nutrient into the interfacial region.*

same mobility parameters and the rescaled ones, respectively. Figure 6 shows the growth of the polymer volume fraction in one-dimensional biofilms according to the MCH model for the same set of values of λ as well as the rescaled one $\lambda \to \lambda/0.19$, respectively. They are qualitatively the same as the results obtained from the CH model, but the transport effect in the MCH model yields weaker polymeric fluxes. For example, the expansion of the biomass predicted by the MCH model with the rescaled mobility is slower than that done by the CH model. In the two MCH models discussed here, the one with the original (nonscaled) mobility parameters clearly delivers weaker polymeric flux, so that the profile of the polymer volume fraction in the majority biofilm-rich region is always higher than those predicted by others. Figure 6 depicts the nutrient concentration calculated from the two models with the same set of mobility parameters as well. The slightly higher nutrient concentration in the case of the MCH solution without rescaling the mobility parameter correlates well with the volume fraction profile at $t = 400$, justifying the fact that the growth is fueled by the supply of the nutrient.

In the one-dimensional situation, the average velocity, the pressure, and the elastic stress tensor components are driven dynamical variables in that their governing equations decouple from the transport equation for ϕ_n and c. We next examine the driven quantities in the one-dimensional models. First, we note that $v_y = 0$ is dictated by the continuity equation. Hence, the polymer network velocity is actually given by (v_x, v_y^e). Figure 7 plots the average velocity component v_x and the excessive velocity component v_y^e. The initial profile of v_x is zero in the biofilm region and nonlinear meeting the prescribed terminal shear speed at $y = 1$. The magnitude of v_x is small in the biofilm region, and all models give comparable predictions. In the solvent region, the CH model gives the largest v_x, while the MCH model of either rescaled or nonscaled mobility parameters is comparable at small mobility and distinct at a larger mobility value. The magnitude of v_x is much smaller in the biofilm region, indicating a lack of spatial motion in the biofilm despite the weak shear. The excessive velocity v_y^e is zero in the solvent region and nonzero in biofilms at $t = 400$. The behavior of v_y^e in the range of small mobility parameters is qualitatively the same. However, the velocity predicted using the MCH model differs from that of the CH model as the mobility increases. In the latter case, the velocities in y predicted by the MCH are all positive, indicating a slight transient growth in the volume fraction at $t = 400$. The negative velocity in the CH model prediction indicates a transient decay of the polymer volume fraction. In all cases, the difference as well as the magnitudes are rather small (on the order of $O(10^{-4})$).

Figure 7 also depicts the normal stress component $\phi_n \tau_{yy}$, where the JSN model with $a = 0.92$ is used. The normal stress components predicted by the three models are similar qualitatively at small mobility parameter $\lambda = 10^{-10}$, where the stress component exhibits a peak in the middle of the biofilm region and a negative value at the biofilm-solvent interface. The same qualitative behavior can be described for the pressure. At higher mobility values, the stress component and the pressure calculated from the CH model yield the largest stress fluctuation in a neighborhood of the interface. The stress obtained from the MCH model with rescaled and nonscaled mobility parameters shows larger numerical values in the biofilm region and smaller fluctuations across the interface. The CH model predicts a stress and pressure undershoot followed by an overshoot in the biofilm region near the interface. Since the transport equation for the polymer volume fraction impacts the polymer network velocity, which in turn drives the polymer elastic stress as well as the pressure, the drastically different behavior is another manifestation of the velocity difference in v_y^e near the interface.

We have contrasted the prediction of the CH model with that of the MCH model. One question remains: which one is better suited for modeling biofilms numerically? In the CH model, the polymeric flux is completely controlled by the variation of the free energy density, while it depends on both the polymer volume fraction and the free energy density variation in the MCH model. Figure 8 depicts the computed profile of the polymer volume fraction in one-dimensional biofilms with a higher nutrient concentration $c^\star = 0.2$ at $y = 1$ and two different mobility parameters. The higher concentration tends to speed up the polymer network expansion and growth across the entire domain. For the CH model, when t is small, we observe that the polymer



(a) $\lambda = 10^{-10}$.

(b) $\lambda = 10^{-8}$.

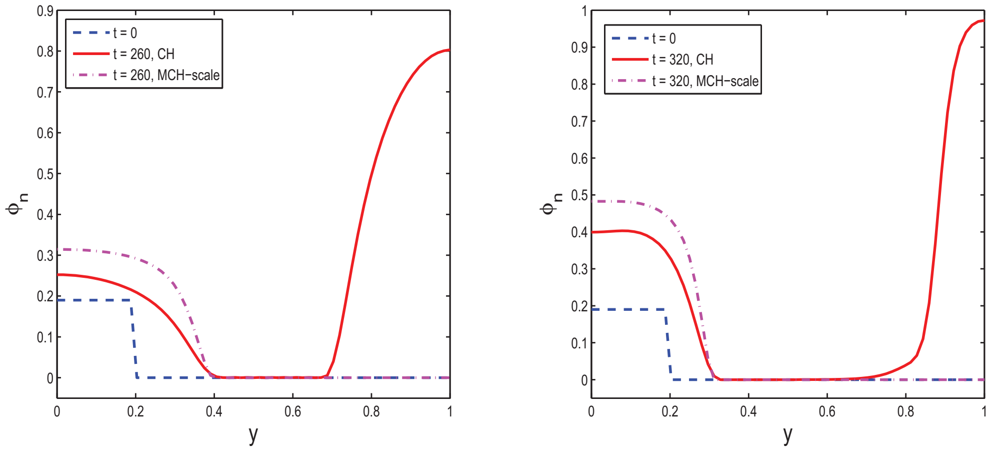
(c) $\lambda = 10^{-10}$.

(d) $\lambda = 10^{-8}$.

(e) $\lambda = 10^{-10}$.

(f) $\lambda = 10^{-8}$.

FIG. 7. *The profile of the average velocity* $\mathbf{v}_x$, v_y^e *and the elastic stress component* $\phi_n \tau_{yy}$ *in the one-dimensional biofilm and solvent mixture.*



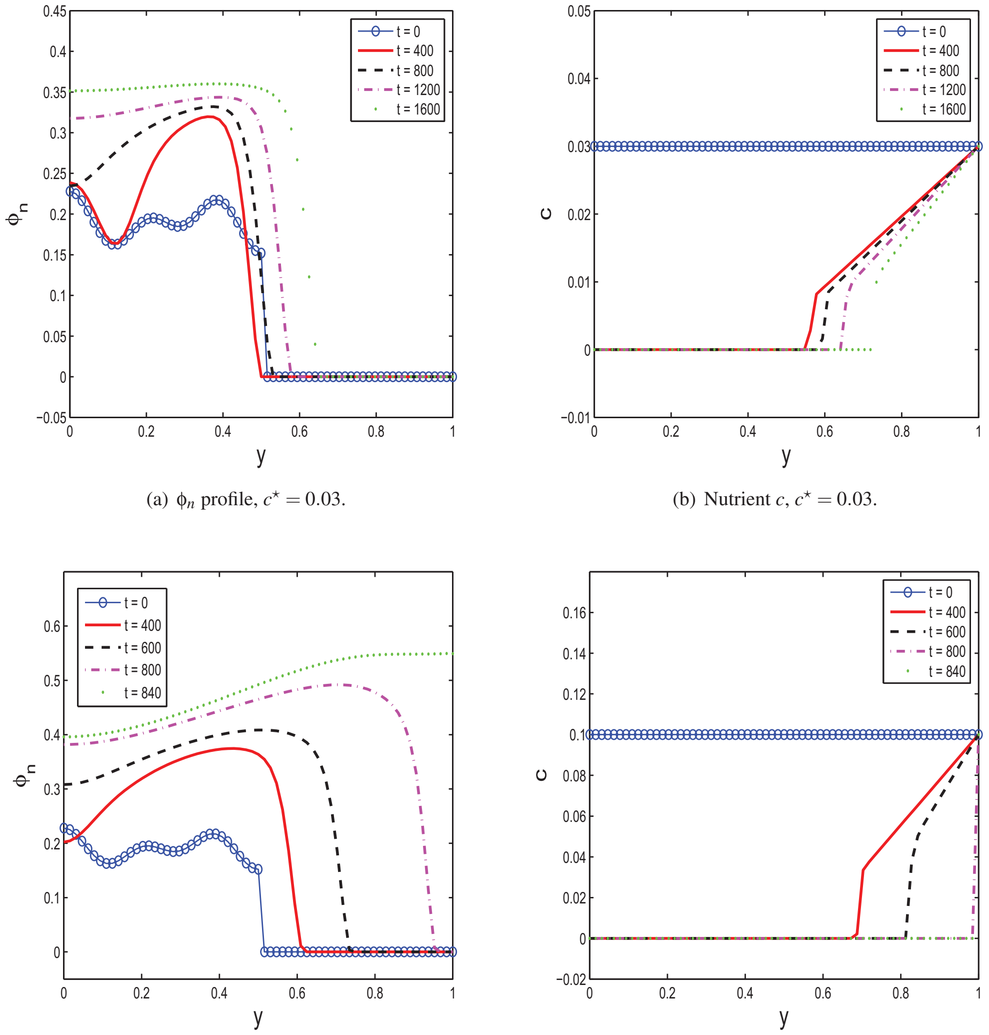
(a) $\lambda = 10^{-9}$.

(b) $\lambda = 10^{-10}$.

FIG. 8. *Growth of the polymer volume fraction in one-dimensional biofilms with a higher nutrient concentration $c^\star = 0.2$. For the CH model, the numerically generated artificial growth at the top boundary disqualifies the model when polymer production is present. The MCH model renders a physically correct prediction in the nutrient-rich solvent, making it our choice of models for studying fluid mixtures.*

network grows due to the production term and expands to the solvent region due to the excessive polymeric flux. As t increases, ϕ_n becomes nonzero at $y = 1$ due to the numerical dissipation and the truncation errors. When ϕ_n becomes nonzero at $y = 1$, an exponential growth ensues due to growth rate g_n in the governing equation for ϕ_n and soon reaches 1, causing our computations to break down. The value of ϕ_n at $y = 1$ reaches 1 faster for $\lambda = 10^{-9}$ (shortly after $t = 260$) than for $\lambda = 10^{-10}$ (shortly after $t = 320$). This numerical evidence demonstrates the limitation of the CH model in modeling the polymer production numerically. The MCH model, on the other hand, does not suffer the unphysically numerical growth of ϕ_n at $y = 1$, since the polymeric flux near $y = 1$ vanishes due to $\phi_n = 0$ in the pure solvent region. Numerically, the zero polymeric flux condition in the solvent region is much easier to maintain in the MCH model before the growth reaches the boundary than in the CH model. We also notice that ϕ_n grows faster near the original interface in the MCH model than in the CH model. This is because once ϕ_n starts to grow at $y = 1$ in the CH model, the nutrient is consumed there quickly, which in turn reduces the amount of the nutrient being diffused to the original interface and thus reduces the polymer production rate. Physically, the MCH model is based on a better assumption on the polymeric flux. The above numerical result hence supports that the MCH model is more appropriate for modeling the transport of the polymer network for its accurate modeling of the transport of the polymeric flux than the CH model.

Finally, we investigate the expansion and growth of an inhomogeneous biofilm initially located at one side of the domain, shown in Figure 9, using the MCH model. When the initial profile contains unstable long wave modes, the dominating growth occurs in the biofilm region with significant coarsening and little expansion into the solvent region initially. In this calculation, we solve the governing system of equations at $\chi = 0.65$ using the MCH model for an extended period of time. Figure 9 depicts the expansion and growth of biofilms from perturbed initial data calculated by the MCH (with the rescaled mobility) model at two different values of nutrient-supply



(a) ϕ_n profile, $c^\star = 0.03$.

(b) Nutrient c, $c^\star = 0.03$.

(c) ϕ_n profile, $c^\star = 0.1$.

(d) Nutrient c, $c^\star = 0.1$.

Fig. 9. *The polymer volume fraction ϕ_n and nutrient concentration c computed by the MCH model with a perturbed initial data with growth for two different $c^\star$, simulated for a sufficiently long time. The parameter values are $\chi = 0.65$, $c^\star = 0.03, 0.1$, $\lambda = 10^{-9}$, $\phi_0(y) = 0.19 + 0.019[\cos(2\pi \cdot 3 \cdot y) + \cos(2\pi \cdot 5 \cdot y)]$ for $y \leq 0.5$, $\phi_0(y) = 0$ for $y > 0.5$. After an initial pulling back, the biofilm expands into the solvent region as long as there is a continuous supply of nutrient.*

boundary value $c^\star$ and for sufficiently long time. Figures 9(a) and (b) show the profile of ϕ_n and c for $c^\star = 0.03$ at five different times. Since $\chi = 0.65$, we know that there are some long wave unstable modes from the linear stability analysis to fuel the expansion and growth of the biofilm. For relatively short time ($t \leq 400$), we observe that the long wave to the left of the interface grows, and the polymer volume fraction profile undergoes a sharp transition near the interface, pulling the polymer network to the biofilm-rich region relative to the initial profile. An intuitive explanation for this is that due to the weaker dissipation in the model, the rapid growth of the polymer network and coarsening in the biofilm draw the polymers near the interface into the polymer-rich region, a consequence of the long wave instability.

The pulling back phenomenon is clearly tied to the coarsening, because the nutrient concentration at $t = 400$ is nearly zero in the biofilm region shown in the figure. As time goes by, we observe that the profile of the polymer volume fraction in the biofilm tends to level off, or coarsening ceases, so that the growth of the polymer network becomes more uniform away from the biofilm-solvent interface and the interface starts to expand into the solvent. From the bulk free energy density $\hat{f}(\phi_n)$, we can see that the bulk contribution to the polymer network flux decreases as the volume fraction ϕ_n continuously grows. At a lower polymer volume fraction, the expansion of the biofilm is facilitated by the bulk free energy along with the conformational free energy tied to the curvature of the interface profile of ϕ_n. As the polymer volume fraction exceeds a critical value though (zero of $\frac{\partial^2 \hat{f}}{\partial \phi_n^2} = 0$), the driving force behind the expansion is due purely to the curvature effect.

We also examine the nutrient distribution during the above-mentioned process. We notice that the nutrient tends to be depleted within the biofilm as the polymer network tends to reach a uniform distribution; however, the nutrient supply is sufficient at the biofilm-solvent interface which fuels the expansion and growth of the polymer network continuously outwards. This explains the dynamics of the polymer network expansion in biofilms for long times. Figures 9(c) and (d) depict the results for a higher nutrient concentration at $c^\star = 0.1$. They are qualitatively the same as the case of $c^\star = 0.03$, except that the dynamics take place at a much faster pace here.

7. Conclusions. In this paper, we present a phase field theory modeling biofilm and solvent mixtures as incompressible complex fluids. In this one-fluid two-component theory, the extracellular polymeric substance (EPS) along with the bacteria is treated as one effective viscous or viscoelastic component, and the nutrient and the solvent are treated as the other effective viscous component. The growth of the effective polymer network component is modeled by a saturated growth, while the nutrient consumption is approximated by a linear decay. Three constitutive models for the mixture are proposed: extended Newtonian, rubber elastic gel, and viscoelastic model. That the mixture in the bulk is incompressible leads to a divergence-free averaged velocity field. The interpenetrating between the two effective components is measured by the excessive velocities accounted for by the Flory–Huggins polymer mixing dynamics. Surface tension between the pure solvent section of the solvent fluid and the biofilm is naturally built in through a nonlocal entropic mixing free energy density. The Cahn–Hilliard dynamics coupled with the Flory–Huggins mixing is investigated with respect to various mobility parameters. Modified Cahn–Hilliard dynamical transport is shown to be more appropriate for the biofilm expansion and growth, which can effectively eliminate the unwanted and unphysical growth in the solvent region due to the numerical error and dissipation.

There are a limited number of results that can be used to validate the model presently. One of the results used in a few reports [8, 13, 25] is that the flat biofilm-fluid interfaces are unstable for a finite interval of perturbation modes, with a single maximally unstable mode. Both the linear analysis and the nonlinear simulations of the present model confirm these predictions.

The advantage of modeling biofilms using a multicomponent material includes robust treatment of the physics and interacting dynamics among the components. Meanwhile, deriving a model consisting of a single fluid eliminates several potential difficulties associated with the coupled biofilm-bulk fluid flow like velocity, boundary conditions, etc. In particular, the interface conditions are dramatically simplified, since the interface is not separated from the rest of the system. In addition, influent and effluent boundary conditions are natural in the single fluid case. The present treatment also provides a framework in which various constitutive relations for each constituent can be investigated in conjunction with the motion of the bulk fluid. Both of these are important in order to address dispersal, detachment, and sloughing events which have substantial impact in industrial and medical settings of the biofilm.

REFERENCES

[1] E. Alpkvist and I. Klapper, *A multidimensional multispecies continuum model for heterogeneous biofilm development*, Bull. Math. Biol., 69 (2007), pp. 765–789.

[2] A. N. Beris and B. Edwards, *Thermodynamics of Flowing Systems*, Oxford Science Publications, New York, 1994.

[3] R. B. Bird, R. C. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, Vols. 1 & 2, John Wiley and Sons, New York, 1987.

[4] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system.* I: *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[5] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system*—III: *Nucleation in a 2-component incompressible fluid*, J. Chem. Phys., 31, (1959), pp. 688–699.

[6] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, Cambridge University Press, Cambridge, UK, 1995.

[7] N. G. Cogan and J. P. Keener, *Channel formation in gels*, SIAM J. Appl. Math., 65 (2005), pp. 1839–1854.

[8] N. Cogan and J. Keener, *The role of biofilm matrix in structural development*, Math. Med. Biol., 21 (2004), pp. 147–166.

[9] J. W. Costerton, Z. Lewandowski, D. E. Caldwell, D. R. Korber, and H. M. Lappin-Scott, *Microbial biofilms*, Annu. Rev. Microbiol., 49 (1995), paper 711745.

[10] B. Costerton, *Medical Biofilm Microbiology: The Role of Microbial Biofilms in Disease, Chronic Infections, and Medical Device Failure*, CD-ROM, Montana State University, Bozeman, MT, 2003.

[11] M. E. Davey and G. A. O'Toole, *Microbial biofilms: From ecology to molecular genetics*, Microbiol. Molec. Biol. Rev., 64 (2000), pp. 847–867.

[12] E. De Lancey Pulcini, *Bacterial biofilms: A review of current research*, Nephrologie, 22 (2001), pp. 439–441.

[13] J. Dockery and I. Klapper, *Finger formation in biofilm layers*, SIAM J. Appl. Math., 62 (2001), pp. 853–869.

[14] M. Doi, *Introduction to Polymer Physics*, Oxford Science Publications, Oxford, UK, 1995.

[15] P. J. Flory, *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, NY, 1953.

[16] D. J. Hassett, P. A. Limbach, R. F. Hennigan, K. E. Klose, R. E. Hancock, M. D. Platt, and D. F. Hunt, *Bacterial biofilms of importance to medicine and bioterrorism: Proteomic techniques to identify novel vaccine components and drug targets*, Expert Opin. Biol. Ther., 3 (2003), pp. 1201–1207.

[17] C. A. A. Lima, R. Ribeiro, E. Foresti, and M. Zaiat, *Morphological Study of Biomass During the Start-Up Period of a Fixed-Bed Anaerobic Reactor Treating Domestic Sewage*, Brazilian Archives of Biology and Technology, 48 (2004), pp. 841–849.

[18] I. Klapper, *Effect of heterogeneous structure in mechanically unstressed biofilms on overall growth*, Bull. Math. Biol., 66 (2004), pp. 809–824.

[19] I Klapper, C. J. Rupp, R. Cargo, B. Purvedorj, and P. Stoodley, *Viscoelastic fluid description of bacterial biofilm material properties*, Biotech. Bioeng., 80 (2002), pp. 289–296.

[20] I. Klapper and J. Dockery, *Role of cohesion in the material description of biofilms*, Phys. Rev. E, 74 (2006), paper 031902.

[21] C. S. Laspidou and B. E. Rittmann, *Modeling biofilm complexity by including active and inert biomass and extracellular polymeric substances*, Biofilm, 1 (2004), pp. 285–291.

[22] J. Lowengrub and L. Truskinovsky, *Quasi-incompressible Cahn–Hilliard fluids and topological transitions*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 2617–2654.

[23] S. T. Milner, *Dynamical theory of concentration fluctuations in polymer solutions under shear*, Phys. Rev. E, 48 (1993), pp. 3674–3691.

[24] G. O'Toole, H. B. Kaplan, and R. Kolter, *Biofilm formation as microbial development*, Ann. Rev. Microbiol., 54 (2000), pp. 49–79.

[25] C. Picioreanu, M. van Loosdrecht, and J. Heijnen, *Mathematical modeling of biofilm structure with a hybrid differential-discrete cellular automaton approach*, Biotech. Bioeng., 58 (1998), paper 101116.

[26] C. Picioreanu, M. van Loosdrecht, and J. Heijnen, *Multidimensional modelling of biofilm structure*, Biotech. Microbial Biosystems: New Frontiers, Proceedings of the 8th International Symposium on Microbial Ecology, C. R. Bell, M. Brylinsky, and P. Johnson-Green, eds., Atlantic Canada Society for Microbial Ecology, Halifax, NS, Canada, 1999, pp. 1–12.

[27] C. Picioreanu, M. J.-U. Kreft, and M. van Loosdrecht, *Particle-based multidimensional multispecies biofilm models*, Appl. Environ. Microbiol., May (2004), pp. 3024–3040.

[28] C. Picioreanu, J. B. Xavier, and M. van Loosdrecht, *Advances in mathematical modeling of biofilm structure*, Biofilm, 1 (2004), pp. 337–349.

[29] H. Tanaka, *Viscoelastic model of phase separation*, Phys. Rev. E, 56 (1997), pp. 4451–4462.

[30] O. Wanner and W. Gujer, *A multispecies biofilm model*, Biotech. Bioeng., 28 (1986), pp. 314–328.

[31] C. Wolgemuth, E. Hoiczyk, D. Kaiser, and G. Oster, *How Myxobacteria glide*, Current Biol., 12 (2002), pp. 369–377.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 670–689

RECURSIVE DISPERSION RELATIONS IN ONE-DIMENSIONAL PERIODIC ELASTIC MEDIA[*]

ANI P. VELO[†], GEORGE A. GAZONAS[‡], ERWIN BRUDER[†§], AND NANCY RODRIGUEZ[†]

Abstract. A frequency bandgap is a range of wave frequencies that are prohibited from passing through a medium. The dispersion relation, which links the frequency to the wave number, enables us to illustrate the bandgaps. In [E. H. Lee, "A survey of variational methods for elastic wave propagation analysis in composites with periodic structures," in Dynamics of Composite Materials, E. H. Lee, ed., ASME, New York, 1972, pp. 122–138] and [E. H. Lee and W. H. Yang, *SIAM J. Appl. Math.*, 25 (1973), pp. 492–499] the dispersion relation was studied theoretically for the one-dimensional periodic structure made of two materials arranged symmetrically with respect to the center of the cell. Their dispersion relation formulas can be similarly extended to a multilayered symmetric cell configuration, but not to a general (nonsymmetric) cell configuration. The general model was considered in [M. Shen and W. Cao, *J. Phys. D*, 33 (2000), pp. 1150–1154], where each unit cell of the periodic layered structure contains several sublayers of arbitrary lengths and materials. Using the transfer matrix method, the dispersion relation was successfully derived, involving very lengthy explicit formulas. In this paper, we generalize the work of Lee and Yang and develop recursive dispersion relation formulas for a general cell configuration. The recursive formulas are easy to implement and, through several numerical experiments, successfully corroborate the results of Shen and Cao.

Key words. dispersion relation, recursive formulas, wave propagation, Floquet theory, periodic layered media

AMS subject classifications. 35C05, 35R05

DOI. 10.1137/070692595

1. Introduction.

The existence of bandgaps in one-dimensional periodic elastic media appears to have been first established by Lord Rayleigh [17], and a good review of the early work on wave propagation in periodic elastic media can be found in Brillouin's classic text [2]; a more recent comprehensive review that outlines the development of the "band theory" for electrons, photons, and phonons can be found in Kushwaha [10]. A variety of technological applications has been suggested for phononic bandgap materials which include transducers, acoustic filters, or barriers for noise reduction, and even as a means for mitigating the effects of seismic surface waves. The study of phonons and phononic bandgaps associated with elastic wave propagation in periodic elastic media have also been used to study quantum field effects such as tunneling phenomena [24].

[*] Received by the editors May 21, 2007; accepted for publication (in revised form) July 17, 2008; published electronically December 5, 2008. This research effort was partially sponsored by the Faculty Research and Engineering Program (FREP), administered by the Battelle Memorial Institute in North Carolina for the U.S. Army Research Laboratory. This work was performed in part by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siap/69-3/69259.html

[†] Department of Mathematics and Computer Science, University of San Diego, 5998 Alcalá Park, San Diego, CA 92110 (avelo@sandiego.edu).

[‡] U.S. Army Research Laboratory, Weapons and Materials Research Directorate, Aberdeen Proving Ground, MD 21005 (gazonas@arl.army.mil).

[§] The École Spéciale Militaire de Saint-Cyr, Coëtquidan, France.

Exact dispersion relations for harmonic waves in an infinite one-dimensional medium consisting of plane parallel alternating layers of two homogeneous isotropic elastic materials were derived in [20]. Their exact dispersion curves compare well with the "effective stiffness" theory they developed for the lowest vibrational modes over a wide range of wave numbers. Using a transfer matrix formalism and bypassing the use of Floquet theory, "pseudo-"stop bands and pass bands were computed in [3] for finite, periodically layered media. They showed that for a finite system containing at least ten cells, the characteristics of the second stop band compare well with that predicted in an infinite medium [11], [12] for a 2-2 composite consisting of ceramic and polymer constituents. This problem was further examined in [7], where it was shown that, in some instances, only one or two unit cells could be sufficient to depict the "frequency bandedness" seen in the infinite medium. Dispersion effects in finite periodic structures, which include viscous damping and the use of genetic algorithms for tailoring their frequency response characteristics, are also considered in [8].

Interestingly, the transfer matrix formalism has been successfully used for some time by geophysicists (e.g., [21], [5], [16], [1]), and for finite, layered Goupillaud-type (equal travel time) media [4], also known as the so-called communication matrix approach [19], [22], [9]. These works are not usually cited in prior work by the "bandgap" community, but are included here to emphasize their importance in providing insight and a framework for the analysis of dispersion effects in periodic elastic media.

Returning our attention once again to infinite media, [11] and [12] study the dispersion relation in an infinite strip of a periodically repeated cell with length or period a. According to their model, the cell is composed of two homogeneous elastic materials: the filler (f) and the matrix material (m). These materials are symmetrically arranged with respect to the center of the unit cell as shown in Figure 1.

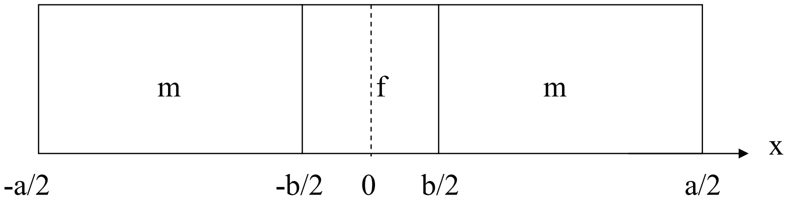


FIG. 1. *Symmetric unit cell made of two materials, used in* [11], [12].

The material density ρ and elastic modulus η are piecewise constant functions, taking constant values with subscripts f and m in the filler and matrix material regions, respectively. The density ρ and the elastic modulus η vary periodically along the strip with position x and period a,

$$\begin{cases} \rho(x+a) = \rho(x), \\ \eta(x+a) = \eta(x). \end{cases}$$

Figure 2 displays the density function within a single (unit) cell, assuming that $\rho_f \leq \rho_m$. Since the cell is periodically repeated in the infinite strip, the density graph shown in Figure 2 is also periodically repeated.

The general model was considered in [18], where each unit cell of the periodic layered structure contains several sublayers of arbitrary lengths and materials. They were

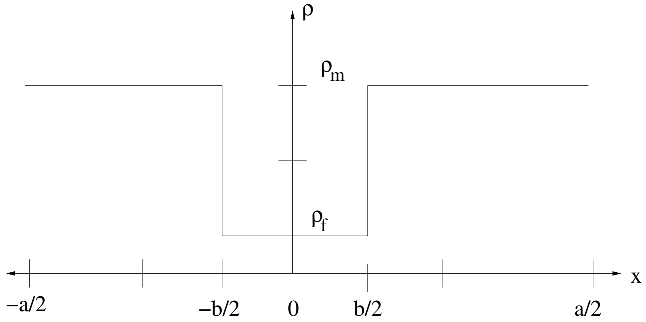


FIG. 2. *Density in a symmetric unit cell made of two materials.*

able to successfully derive the dispersion relation using the transfer matrix method, while involving very long explicit formulas.

In this paper, we generalize the work of [11], [12] and develop recursive dispersion relation formulas for a general cell configuration. The recursive formulas are easy to implement and, through several numerical experiments using Maple [15], successfully corroborate the results of [18].

Unlike the (two-material) symmetric cell configuration studied in [11], [12], the implicit dispersion relation for a general cell configuration appears to be more complex. The process of deriving an implicit recursive dispersion relation involves the construction of basic solutions in the unit cell, which provides more insight on how the properties of such solutions relate to the cell configuration. This is demonstrated through two approaches, which we identify as the central expansion approach and the quasi-symmetric limiting approach.

As shown in [11], [12], we begin with the wave equation with periodic coefficients η and ρ, described by

$$\frac{\partial[\eta\frac{\partial U}{\partial x}]}{\partial x} = \rho\frac{\partial^2 U}{\partial t^2}. \tag{1}$$

Using separation of variables, we assume that the displacement $U(x,t)$ can be expressed as

$$U(x,t) = u(x)\phi(t), \tag{2}$$

and (1) reduces to the second order ordinary differential equation with periodic coefficients,

$$\frac{d}{dx}\left[\eta\frac{du}{dx}\right] + \rho\omega^2 u = 0. \tag{3}$$

According to the Floquet theory, for a fixed ω, the solution $u(x)$ in (3) is of the form

$$u(x) = \upsilon(x)e^{iqx}, \tag{4}$$

where $\upsilon(x)$ is a periodic function with the same period a as the coefficients η and ρ. Due to the quasi-periodic recursive relation that follows from (4), we have

$$u(x+a) = u(x)e^{iqa},$$

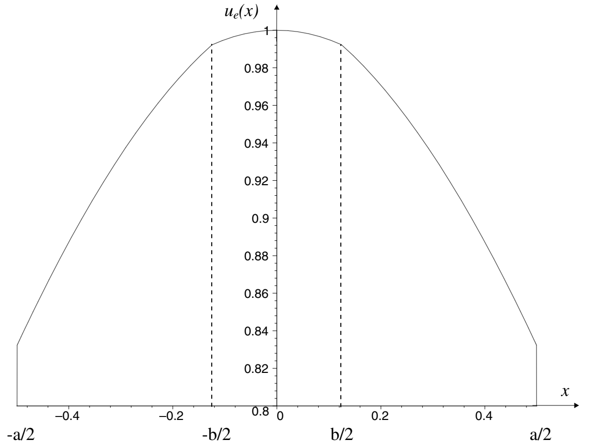


FIG. 3. *Even eigenfunction in a symmetric cell of two materials and three layers.*

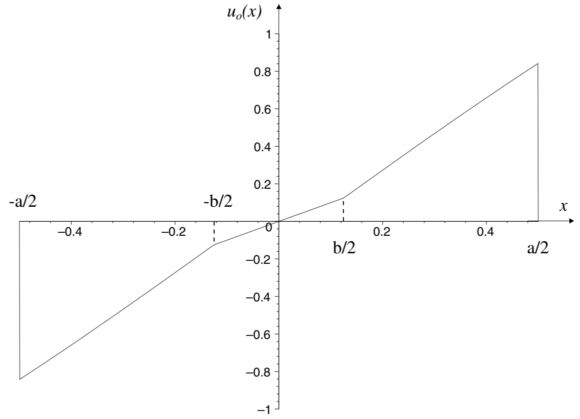


FIG. 4. *Odd eigenfunction in a symmetric cell of two materials and three layers.*

and the problem of finding the solution $u(x)$ along the strip is reduced to the single unit cell $-a/2 \le x \le a/2$, where the following quasi-periodic boundary conditions apply:

$$\left\{ \begin{aligned} u(a/2) &= u(-a/2)e^{iqa}, \\ u'(a/2) &= u'(-a/2)e^{iqa}. \end{aligned} \right. \tag{5}$$

The solution of (3) in the unit cell, subject to the boundary conditions (5), is then expressed as a linear combination of two eigenfunctions/linearly independent solutions of (3). For convenience, the two eigenfunctions are chosen to be even $u_e(x)$ and odd $u_o(x)$ functions along the symmetric cell studied in [11], [12]; see Figures 3 and 4. The solution of (3),

$$u(x) = u_e(x) + Cu_o(x), \tag{6}$$

satisfies the quasi-periodic boundary conditions (5) if

$$C = i\frac{u_e(a/2)}{u_o(a/2)}\tan\frac{qa}{2} \quad \text{and} \quad C = -i\frac{u_e'(a/2)}{u_o'(a/2)}\cot\frac{qa}{2}. \tag{7}$$

According to [11], [12], the compatibility of the two relations above results in the implicit form of the dispersion relation,

$$\boxed{\frac{u_e(a/2)}{u_o(a/2)}\tan\frac{qa}{2} = -\frac{u_e'(a/2)}{u_o'(a/2)}\cot\frac{qa}{2}.} \tag{8}$$

Further simplifications imply the following equivalent forms:

$$\tan^2\left(\frac{qa}{2}\right) = -\frac{u_e'(a/2)u_o(a/2)}{u_o'(a/2)u_e(a/2)}$$

or

$$\boxed{\cos(qa) = \frac{u_o'(a/2)u_e(a/2) + u_e'(a/2)u_o(a/2)}{u_o'(a/2)u_e(a/2) - u_e'(a/2)u_o(a/2)}}. \tag{9}$$

The dispersion relation is then obtained after the construction of the even and odd eigenfunctions $u_e(x)$ and $u_o(x)$ in the unit cell, and their substitution into (8) or (9). As discussed in [11], [12], due to (4), it is only necessary to consider the wave number q limited to the domain $0 \leq q \leq \pi/a$. The dispersion relation graph displays a banded frequency spectrum, comprising bands which transmit Floquet waves and no pass bands which do not.

2. Generalized form of the dispersion relation. In this section, we derive the dispersion relation for a general (unit) cell configuration, made of an arbitrary number of layers and materials. The steps involved are summarized below, and are generalizations of the work of [11], [12] discussed earlier. Through the rest of the paper, the interval of the unit cell with length a is $[-b, d]$. Here $0 < b, d < a$, and $b+d = a$. In the special case of a symmetric cell configuration seen before, $b = d = \frac{a}{2}$.

(i) The application of Floquet's theorem for (3), with ρ and η corresponding to the general (unit) cell configuration, yields two quasi-periodic conditions,

$$\begin{cases} u(x+a) = u(x)e^{iqa}, \\ u'(x+a) = u'(x)e^{iqa}. \end{cases}$$

Evaluating the conditions above at $x = -b$, we obtain the generalization of the boundary conditions (5),

$$\begin{cases} u(d) = u(-b)e^{iqa}, \\ u'(d) = u'(-b)e^{iqa}. \end{cases} \tag{10}$$

(ii) The solution $u = u(x)$ of (3) and (10) may be written in a general form similar to (6) as

$$u = C_1u_1 + C_2u_2. \tag{11}$$

Here $u_1 = u_1(x)$ and $u_2 = u_2(x)$ are two eigenfunctions/linearly independent solutions of (3), to be constructed later, while the constants C_1 and C_2 are unknown. For a symmetric cell configuration, such as the one studied in [11], [12], u_1 and u_2 become even and odd functions.

(iii) As a generalization of (7), after substituting (11) into the quasi-boundary conditions (10), we obtain a linear homogeneous system of equations for the unknowns C_1 and C_2,

$$\left\{\begin{aligned} &(u_1(d) - u_1(-b)e^{iqa})\, C_1 + (u_2(d) - u_2(-b)e^{iqa})\, C_2 = 0, \\ &(u_1'(d) - u_1'(-b)e^{iqa})\, C_1 + (u_2'(d) - u_2'(-b)e^{iqa})\, C_2 = 0. \end{aligned}\right. \tag{12}$$

(iv) Seeking a nontrivial/nonzero solution, we set the determinant of the system (12) to zero, a condition which replaces the compatibility conditions used earlier in (7)–(8) by [11], [12]. This yields the $\omega = \omega(q)$ relation of the form

$$\boxed{\begin{aligned} &(u_1(d) - u_1(-b)e^{iqa})(u_2'(d) - u_2'(-b)e^{iqa}) \\ &= (u_1'(d) - u_1'(-b)e^{iqa})(u_2(d) - u_2(-b)e^{iqa}). \end{aligned}} \tag{13}$$

Equation (13) represents the dispersion relation for a general cell configuration in its implicit complex form.

(v) After a few manipulations of (13), using the property of the Wronskian $W(x)$ discussed below, we obtain the dispersion relation in its implicit real form. The Wronskian $W(x)$ of $u_1(x)$ and $u_2(x)$, two linearly independent solutions of the differential equation (3), is given by

$$W(x) = u_1(x)u_2'(x) - u_1'(x)u_2(x).$$

It directly follows from (3) that

$$\frac{d}{dx}[\eta(x)W(x)] = 0,$$

and therefore,

$$W(x) = \frac{C}{\eta(x)},$$

where C is a constant. Assuming that $\eta(-b) = \eta(d)$, one deduces that $W(-b) = W(d)$. Returning our attention to (13), after multiplying and reorganizing the terms, we obtain

$$W(d) - J(-b, d)e^{iqa} + W(-b)e^{i2qa} = 0, \tag{14}$$

where $J(-b, d) = u_1(-b)u_2'(d) + u_1(d)u_2'(-b) - u_1'(-b)u_2(d) - u_1'(d)u_2(-b)$.

Under the assumption that $\eta(-b) = \eta(d)$, one deduces that $W(-b) = W(d)$ and the relation (14) becomes

$$W(-b)[e^{iqa} + e^{-iqa}] = J(-b, d).$$

From here the dispersion relation is expressed in its implicit real form as

$$2\cos(qa) = \frac{J(-b, d)}{W(-b)},$$

or equivalently

$$\boxed{2\cos(qa) = \frac{u_1(-b)u_2'(d) + u_1(d)u_2'(-b) - u_1'(-b)u_2(d) - u_1'(d)u_2(-b)}{u_1(-b)u_2'(-b) - u_1'(-b)u_2(-b)}.} \tag{15}$$

Further simplifications follow if $u_1(x)$ and $u_2(x)$ are chosen to satisfy $u_1(-b) = u_2'(-b) = 1$ and $u_1'(-b) = u_2(-b) = 0$. Then $W(-b) = 1$ and the implicit dispersion relation,

$$2\cos(qa) = \frac{J(-b,d)}{W(-b)},$$

takes the form

$$2\cos(qa) = u_1(d) + u_2'(d).$$

Similar arguments can be found in [14].

A simplified form of (15) used to express the dispersion relation for the (two-material) symmetric cell configuration was derived in [11], [12] and given by (9). Indeed, when $b = d = \frac{a}{2}$, we obtain

$$(16)\qquad 2\cos(qa) = \frac{u_1(-\frac{a}{2})u_2'(\frac{a}{2}) + u_1(\frac{a}{2})u_2'(-\frac{a}{2}) - u_1'(-\frac{a}{2})u_2(\frac{a}{2}) - u_1'(\frac{a}{2})u_2(-\frac{a}{2})}{u_1(-\frac{a}{2})u_2'(-\frac{a}{2}) - u_1'(-\frac{a}{2})u_2(-\frac{a}{2})}.$$

Due to the symmetry of the cell configuration, u_1 and u_2 become even and odd functions, respectively. This means that

$$\begin{cases} u_1(\frac{a}{2}) = u_1(-\frac{a}{2}), \\ u_1'(\frac{a}{2}) = -u_1'(-\frac{a}{2}) \end{cases} \quad\text{and}\quad \begin{cases} u_2(\frac{a}{2}) = -u_2(-\frac{a}{2}), \\ u_2'(\frac{a}{2}) = u_2'(-\frac{a}{2}). \end{cases}$$

After substituting these relations into (16) and replacing $u_1 = u_e$ and $u_2 = u_o$, we obtain the dispersion relation (9), as seen before in [11], [12].

As a conclusion, relation (15) subject to the boundary requirement $\eta(-b) = \eta(d)$ represents the generalized form of the dispersion relation (9) derived in [11], [12]. This is also confirmed by our numerical experiments with the choice of the unit cell boxed in Figure 5, with border layers occupied by the same material, ensuring the same value for $\eta(x)$ and therefore the same value for $W(x)$ along the border layers.

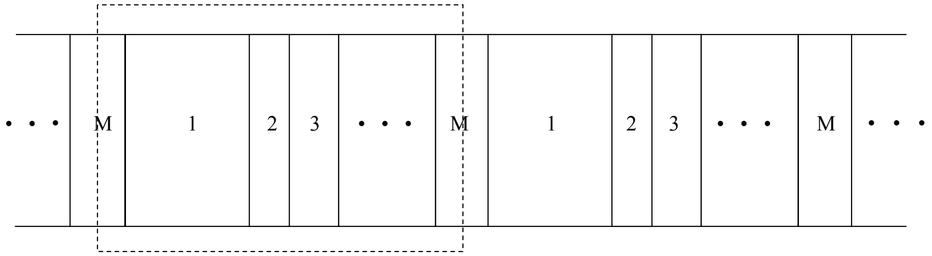


FIG. 5. *Unit cell selection in a periodic medium with a general cell configuration.*

3. Recursive formula of the dispersion relation using the central expansion approach. In order to develop the generalized dispersion relation (15), we need to find two eigenfunctions $u_1(x)$ and $u_2(x)$ of (3) in a unit cell of our choice. The unit cell of choice, used in the central expansion approach, is shown in the last diagram of Figure 6. This is obtained after shifting and renumbering the general cell

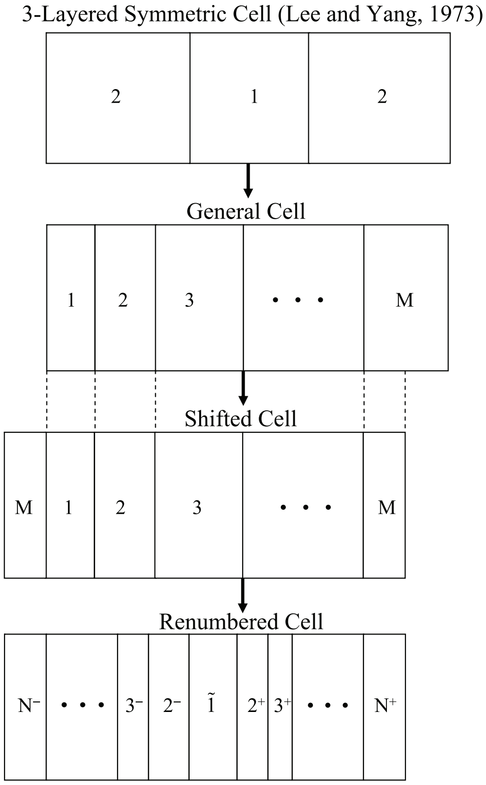


FIG. 6. *The stages of development of the central expansion approach for a given general cell configuration of M-layers. Here $N^+ = N^- = N$, where $N = \lceil \frac{M}{2} \rceil + 1$ or $N = \lceil \frac{M}{2} \rceil$.*

of M-layers. The purpose of the shifting is to have the border layers made of the same material to ensure that $\eta(-b) = \eta(d)$, which implies that $W(-b) = W(d)$ and therefore (15) holds. The purpose of the renumbering is to create a central layer in the cell similar to the three-layered symmetric cell configuration previously studied in [11], [12]. In the renumbered scheme, $N = \lceil \frac{M}{2} \rceil + 1$, unless the cell already has an odd number of layers M and the border layers are made of the same material, in which case $N = \lceil \frac{M}{2} \rceil$. The $(-)$ and $(+)$ superscripts on the renumbered cell diagram of Figure 6 are used for the layers numbered $2, 3, \ldots, N$ to indicate, respectively, their left and right positions with respect to the central layer. The central layer is marked $\tilde{1}$ to distinguish it from the layers on the other diagrams marked 1.

As a result, we may follow the method of [11], [12]. We begin with the solution in the central layer of the cell and expand it to the right and left layers using the continuity conditions of stress and displacement at the layer interfaces. Indeed, in a given layer of the cell, with constant material properties ρ and η and wave speed $c = \sqrt{\eta/\rho}$, the solution of (3) is of the form

$$u(x) = A \cos\left(\frac{\omega}{c} x\right) + B \sin\left(\frac{\omega}{c} x\right). \tag{17}$$

In the notation that follows $\rho_j^{\pm}$, $\eta_j^{\pm}$, and wave speed $c_j^{\pm} = \sqrt{\eta_j^{\pm}/\rho_j^{\pm}}$ indicate the corresponding values for the jth layer, $j = 2, \ldots, N$, located to the right ($(+)$ superscript) or left ($(-)$ superscript) of the central layer with $j = 1$.

As indicated in [11], [12], it is convenient to choose the two eigenfunctions $u_1(x)$ and $u_2(x)$ to be even and odd functions in the central layer,

$$u_1(x) = \cos\left(\frac{\omega}{c_1}x\right), \ -b_1 < x < d_1, \tag{18}$$

$$u_2(x) = \sin\left(\frac{\omega}{c_1}x\right), \ -b_1 < x < d_1. \tag{19}$$

The expansion of these solutions to the right and left layers would be of the form (17), and the eigenfunction $u_1(x)$ along the unit cell $[-b, d]$ can be given as

$$u_1(x) = \begin{cases} A_N^+ \cos\left(\frac{\omega}{c_N^+}(x - d_{N-1})\right) + B_N^+ \sin\left(\frac{\omega}{c_N^+}(x - d_{N-1})\right) & \text{for } d_{N-1} \le x \le d_N, \\ \vdots & \\ A_3^+ \cos\left(\frac{\omega}{c_3^+}(x - d_2)\right) + B_3^+ \sin\left(\frac{\omega}{c_3^+}(x - d_2)\right) & \text{for } d_2 \le x \le d_3, \\ A_2^+ \cos\left(\frac{\omega}{c_2^+}(x - d_1)\right) + B_2^+ \sin\left(\frac{\omega}{c_2^+}(x - d_1)\right) & \text{for } d_1 \le x \le d_2, \\ \cos\left(\frac{\omega}{c_1}x\right) & \text{for } -b_1 \le x \le d_1, \\ A_2^- \cos\left(\frac{\omega}{c_2^-}(x + b_1)\right) + B_2^- \sin\left(\frac{\omega}{c_2^-}(x + b_1)\right) & \text{for } -b_2 \le x \le -b_1, \\ A_3^- \cos\left(\frac{\omega}{c_3^-}(x + b_2)\right) + B_3^- \sin\left(\frac{\omega}{c_3^-}(x + b_2)\right) & \text{for } -b_3 \le x \le -b_2, \\ \vdots & \\ A_N^- \cos\left(\frac{\omega}{c_N^-}(x + b_{N-1})\right) + B_N^- \sin\left(\frac{\omega}{c_N^-}(x + b_{N-1})\right) & \text{for } -b_N \le x \le -b_{N-1}. \end{cases} \tag{20}$$

Using vector notation we denote

$$v_j^\pm = \begin{bmatrix} A_j^\pm \\ B_j^\pm \end{bmatrix}$$

for $j = 1, \ldots, N$.

The coefficients in (20), derived by the continuity conditions at the layer interfaces, are given by the recursive relations

$$v_{j+1}^\pm = M_j^\pm v_j^\pm, \tag{21}$$

where the matrix $M_j^\pm$ and the constant parameters are, respectively, given by

$$M_j^\pm = \begin{bmatrix} \cos\lambda_j^\pm & \pm\sin\lambda_j^\pm \\ \mp p_j^\pm \sin\lambda_j^\pm & p_j^\pm \cos\lambda_j^\pm \end{bmatrix} \tag{22}$$

and

$$\begin{cases} \lambda_j^+ = \frac{\omega}{c_j^+}(d_j - d_{j-1}), \ \lambda_j^- = \frac{\omega}{c_j^-}(b_j - b_{j-1}), \ p_j^\pm = \frac{\eta_j^\pm c_{j+1}^\pm}{\eta_{j+1}^\pm c_{j+1}^\pm}, \\ v_1^\pm = \begin{bmatrix} A_1^\pm \\ B_1^\pm \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ d_0 = 0, \ b_0 = 0 \end{cases} \tag{23}$$

for $j = 1, \dots, N-1$. Here A_n^+ and B_n^+ are the coefficients of the eigenfunction $u_1(x)$ along the nth right layer, while A_n^-, B_n^- are the coefficients along the nth left layer. The layer interfaces are located at $-b = -b_N < -b_{N-1} < \cdots < -b_n < \cdots < -b_1 < d_1 < \cdots < d_n < \cdots < d_{N-1} < d_N = d$. Here $b_1 = d_1$, $b + d = a$ and $b_n > 0$, $d_n > 0$ for $n = 1, 2, \dots, N$. Notice that the eigenfunction $u_1(x)$ is even in the central layer, but it does not necessarily remain even after it expands to the other layers of the cell; see Figure 7. The cell in Figure 7 is composed of five layers. Notice that at the layer interfaces, $u_1(x)$ develops corners, as expected from the stress continuity condition.

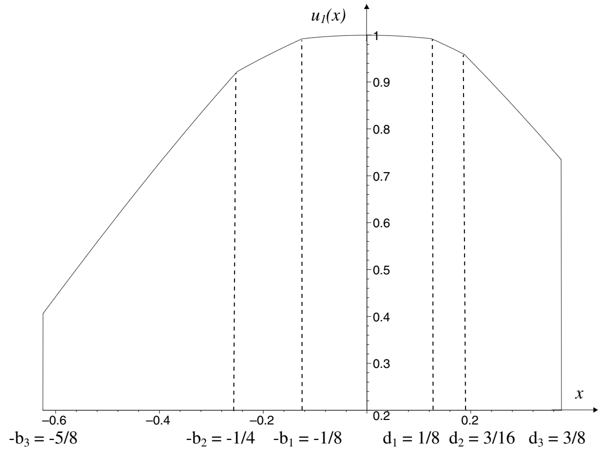


FIG. 7. *Eigenfunction $u_1(x)$ in a nonsymmetric cell of three materials and five layers. Additional parameters involved in* (20) *are* $\omega = c_1 = c_2^\pm = c_3^\pm = 1$, $p_1^\pm = \frac{\eta_1}{\eta_2} = 4$, *and* $p_2^\pm = \frac{\eta_2}{\eta_3} = 2$.

Similarly, we determine the eigenfunction $u_2(x)$ in the unit cell $[-b, d]$ as

$$u_2(x) = \begin{cases} A_N^{*+} \cos\left(\frac{\omega}{c_N^+}(x - d_{N-1})\right) + B_N^{*+} \sin\left(\frac{\omega}{c_N^+}(x - d_{N-1})\right) & \text{for } d_{N-1} \le x \le d_N, \\ \vdots & \\ A_3^{*+} \cos\left(\frac{\omega}{c_3^+}(x - d_2)\right) + B_3^{*+} \sin\left(\frac{\omega}{c_3^+}(x - d_2)\right) & \text{for } d_2 \le x \le d_3, \\ A_2^{*+} \cos\left(\frac{\omega}{c_2^+}(x - d_1)\right) + B_2^{*+} \sin\left(\frac{\omega}{c_2^+}(x - d_1)\right) & \text{for } d_1 \le x \le d_2, \\ \sin\left(\frac{\omega}{c_1}x\right) & \text{for } -b_1 \le x \le d_1, \\ A_2^{*-} \cos\left(\frac{\omega}{c_2^-}(x + b_1)\right) + B_2^{*-} \sin\left(\frac{\omega}{c_2^-}(x + b_1)\right) & \text{for } -b_2 \le x \le -b_1, \\ A_3^{*-} \cos\left(\frac{\omega}{c_3^-}(x + b_2)\right) + B_3^{*-} \sin\left(\frac{\omega}{c_3^-}(x + b_2)\right) & \text{for } -b_3 \le x \le -b_2, \\ \vdots & \\ A_N^{-} \cos\left(\frac{\omega}{c_N^-}(x + b_{M-1})\right) + B_N^{-} \sin\left(\frac{\omega}{c_N^-}(x + b_{N-1})\right) & \text{for } -b_N \le x \le -b_{N-1}. \end{cases}$$

(24)

Using vector notation we denote $v_j^{*\pm} = \left[\begin{smallmatrix} A_j^{*\pm} \\ B_j^{*\pm} \end{smallmatrix}\right]$ for $j = 1, \dots, N$. The coefficients of (24) are defined by the same recursive relations described in (21),

$$v_{j+1}^{*\pm} = M_j^{\pm} v_j^{*\pm}, \tag{25}$$

with

$$v_1^{*\pm} = \begin{bmatrix} A_1^{*\pm} \\ B_1^{*\pm} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{26}$$

The matrix $M_j^{\pm}$ and the constant parameters d_0, b_0, $\lambda_j^{\pm}$, and $p_j^{\pm}$ for $j = 1, \ldots, N-1$ are given as before in (22)–(23).

Here A_n^{*+} and B_n^{*+} are the coefficients of the eigenfunction $u_2(x)$ along the nth right layer, while A_n^{*-}, B_n^{*-} are the coefficients along the nth left layer. Notice that $u_2(x)$ is odd in the central layer, but it does not necessarily remain odd after it expands to the other layers of the cell; see Figure 8. The cell in Figure 7 is composed of five layers. Notice that at the layer interfaces, $u_2(x)$ develops corners, as expected from the continuity of the stress condition.

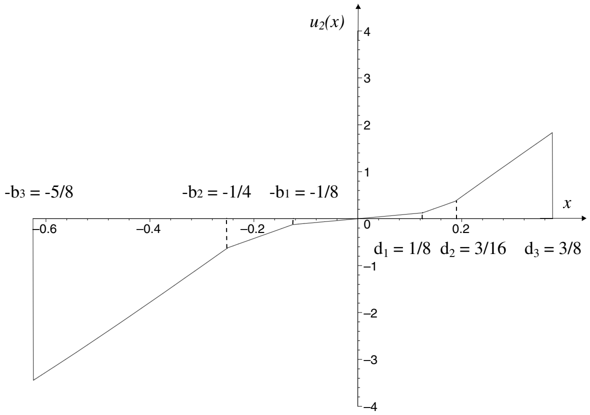


FIG. 8. *Eigenfunction $u_2(x)$ in a nonsymmetric cell of three materials and five layers. Additional parameters involved in* (24) *are $\omega = c_1 = c_2^{\pm} = c_3^{\pm} = 1$, $p_1^{\pm} = \frac{\eta_1}{\eta_2} = 4$, and $p_2^{\pm} = \frac{\eta_2}{\eta_3} = 2$.*

After evaluating $u_1(x)$ and $u_2(x)$ at the boundaries of the cell located at $x = -b$ and $x = d$, and then substituting these expressions into the general implicit formula of the dispersion relation (15), we obtain the following expressions for the numerator and the denominator, respectively:

$$\begin{aligned} &u_1(-b)u_2'(d) + u_1(d)u_2'(-b) - u_2(d)u_1'(-b) - u_2(-b)u_1'(d) \\ &\quad = \frac{\omega}{c_N}(A_N^- B_N^{*+} + A_N^+ B_N^{*-} - A_N^{*+} B_N^- - A_N^{*-} B_N^+)\cos(\lambda_N^+ + \lambda_N^-) \\ &\quad + \frac{\omega}{c_N}(A_N^+ A_N^{*-} - B_N^{*+} B_N^- - A_N^{*+} A_N^{*-} + B_N^{*-} B_N^+)\sin(\lambda_N^+ + \lambda_N^-) \end{aligned} \tag{27}$$

and

$$u_1(-b)u_2'(-b) + u_2(-b)u_1'(-b) = \frac{\omega}{c_N}(A_N^- B_N^{*-} - A_N^{*-} B_N^-). \tag{28}$$

Recall that our leftmost layer is made of the same material as our rightmost layer, and hence $c_N^- = c_N^+ = c_N$.

Substituting (27) and (28) into the general implicit dispersion relation (15), we derive the *recursive formula for the dispersion relation*,

$$\boxed{\begin{aligned}2\cos(qa) = {} & \frac{(A_N^{+}B_N^{*-}+A_N^{-}B_N^{*+}-A_N^{*+}B_N^{-}-A_N^{*-}B_N^{+})\cos(\lambda_N^{+}+\lambda_N^{-})}{A_N^{-}B_N^{*-}-A_N^{*-}B_N^{-}} \\ & + \frac{(A_N^{+}A_N^{*-}-A_N^{-}A_N^{*+}+B_N^{+}B_N^{*-}-B_N^{-}B_N^{*+})\sin(\lambda_N^{+}+\lambda_N^{-})}{A_N^{-}B_N^{*-}-A_N^{*-}B_N^{-}},\end{aligned}} \tag{29}$$

where the coefficients are obtained through the recursive relations involving (21)–(23), (25)–(26), while $\lambda_N^- = \frac{\omega}{c_N^-}(b_N - b_{N-1})$ and $\lambda_N^+ = \frac{\omega}{c_N^+}(d_N - d_{N-1})$. In the numerical experiments included in section 5, the recursive relation generating the coefficients $A_N^{\pm}, B_N^{\pm}, A_N^{*\pm}, B_N^{*\pm}$ of (29) is programmed in Maple [15]. The dispersion relation graph is then obtained using the *implicitplot* Maple command with the wave number q limited to the domain $0 \le q \le \pi/a$.

4. Recursive formula of the dispersion relation using the quasi-symmetric limiting approach. This alternative approach involves the quasi-symmetric cell configuration shown in Figure 9, which preserves the symmetric arrangement of the materials around the filler/central layer, and allows for layers of the same material to have differing lengths.

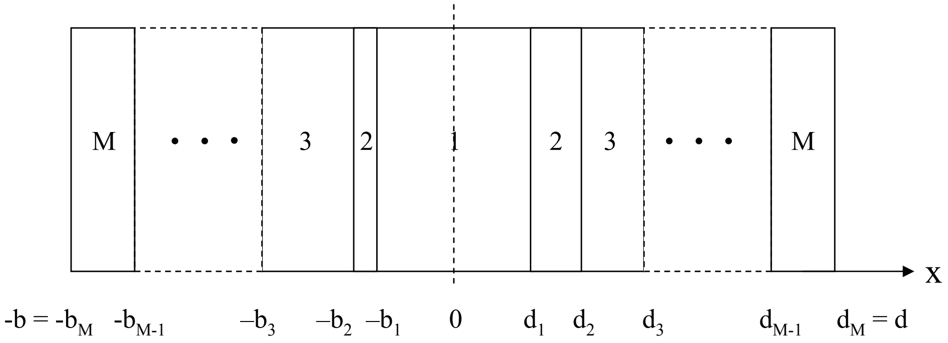


FIG. 9. *Quasi-symmetric configuration of a unit cell made of M materials.*

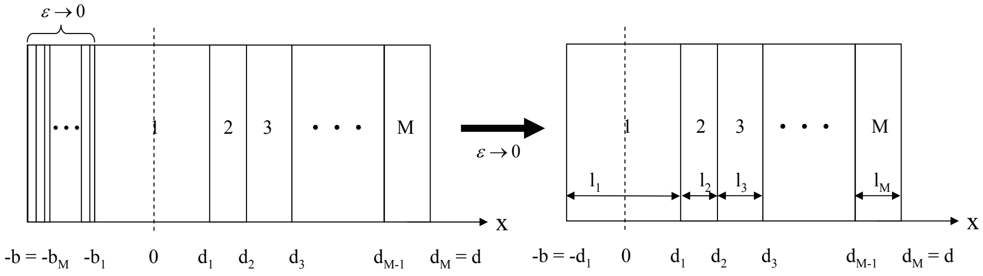


FIG. 10. *Obtaining the general cell configuration as a limiting case of the quasi-symmetric cell configuration.*

The quasi-symmetric limiting approach differs from the central expansion approach in the way the dispersion relation is derived for a general cell configuration with a non-quasi-symmetric configuration. The quasi-symmetric limiting approach views the general cell configuration as a limiting case of the quasi-symmetric configuration when the thickness of the left layer(s) approaches zero; see Figure 10. In the limiting procedure, it is essential to start with a quasi-symmetric and nonsymmetric cell configuration. The same limiting procedure applied to a symmetric cell configuration would fail to produce any results, because in a symmetric cell configuration, as the thicknesses of the left layers approach zero, so do the thicknesses of the corresponding right layers. Thus, the quasi-symmetric nonsymmetric cell configuration is essential, as it allows the layer thicknesses to the left and right of the central layer to be arbitrary, and hence independent. This approach also highlights the fact that the dispersion formula obtained in [11], [12] works for a general symmetric cell configuration but fails once the cell configuration becomes quasi-symmetric. In summary, the quasi-symmetric configuration is a critical configuration to work with, because from there one can recover the dispersion relation for the general cell configuration, something that cannot be achieved from a symmetric configuration. The approach described above is summarized in Figure 11 and discussed in more detail below.

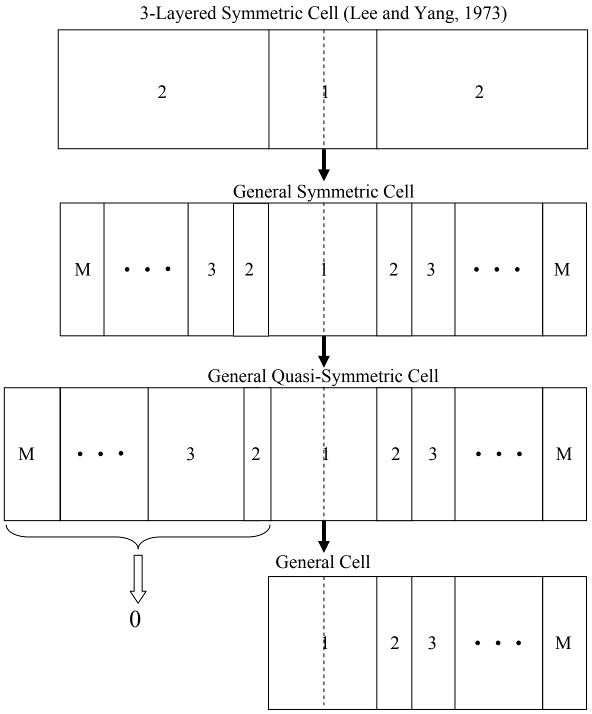


FIG. 11. *The stages of development of the quasi-symmetric limiting approach for a general cell configuration of M-layers.*

The recursive dispersion formula for the quasi-symmetric cell configuration can be derived similarly to that described in the previous section, under the simplifications that $\eta_j^- = \eta_j^+ = \eta_j$, $c_{j+1}^- = c_{j+1}^+ = c_{j+1}$, $p_j^- = p_j^+ = p_j$, and $N = M$ for $j = 1, 2, \ldots, M-1$. With these simplifications, the recursive dispersion formula (29) becomes

$$\boxed{\begin{aligned} 2\cos(qa) = {} & \frac{(A_M^+ B_M^{*-} + A_M^- B_M^{*+} - A_M^{*+} B_M^- - A_M^{*-} B_M^+)\cos(\lambda_M^+ + \lambda_M^-)}{A_M^- B_M^{*-} - A_M^{*-} B_M^-} \\ & + \frac{(A_M^+ A_M^{*-} - A_M^- A_M^{*+} + B_M^+ B_M^{*-} - B_M^- B_M^{*+})\sin(\lambda_M^+ + \lambda_M^-)}{A_M^- B_M^{*-} - A_M^{*-} B_M^-}. \end{aligned}} \tag{30}$$

The coefficients are obtained through the recursive relations involving (21)–(23) and (25)–(26). Notice that $\lambda_M^- = \frac{\omega}{c_M}(b_M - b_{M-1})$, $\lambda_M^+ = \frac{\omega}{c_M}(d_M - d_{M-1})$, while $\lambda_j^- = \frac{\omega}{c_j}(b_j - b_{j-1})$, $\lambda_j^+ = \frac{\omega}{c_j}(d_j - d_{j-1})$, $p_j = \frac{\eta_j c_{j+1}}{\eta_{j+1} c_{j+1}}$ for $j = 1, \ldots, M-1$.

As for the general cell configuration, while the coefficients of (30) with (+) superscript corresponding to the right-hand side layers remain intact, the coefficients of the left-hand side layers with (−) superscript will simplify due to the fact that the left-hand side layers vanish; see Figure 10. From (21)–(23), it follows that the coefficients with (−) superscript related to the eigenfunction $u_1(x)$ along the left layers are given by

$$(31)\qquad \begin{cases} A_{j+1}^- = A_j^- \cos(\lambda_j^-) - B_j^- \sin(\lambda_j^-), \\ B_{j+1}^- = p_j(B_j^- \cos(\lambda_j^-) + A_j^- \sin(\lambda_j^-)), \\ A_1^- = 1,\ B_1^- = 0,\ j = M-1, \ldots, 1. \end{cases}$$

In the limiting case, as $(b_j - b_{j-1})$ approaches zero, so does $\lambda_j = \frac{\omega}{c_j}(b_j - b_{j-1})$ for $j = M, \ldots, 2$. As a result, the recursive relation (31) becomes

$$(32)\quad \begin{cases} A_M^- = A_{M-1}^- = \cdots = A_j^- = \cdots = A_2^-, \\ B_M^- = p_{M-1}B_{M-1}^- = \cdots = p_{M-1}p_{M-2} \cdot p_j B_j^- = \cdots = p_{M-1}p_{M-2} \cdots p_2 B_2^-. \end{cases}$$

Furthermore, considering that $p_j = \frac{\eta_j c_{j+1}}{\eta_{j+1} c_j}$, $\lambda_1^+ = \lambda_1^- = \lambda_1$ (central layer), $A_1^- = 1$, and $B_1^- = 0$, the relations (32) become

$$(33)\quad \begin{cases} A_M^- = A_{M-1}^- = \cdots = A_j^- = \cdots = A_2^- = \cos(\lambda_1), \\ B_M^- = \dfrac{\eta_{M-1} c_M}{\eta_M c_{M-1}} B_{M-1}^- = \cdots = \dfrac{\eta_j c_M}{\eta_M c_j} B_j^- = \cdots = \dfrac{\eta_2 c_M}{\eta_M c_2} B_2^- = \dfrac{\eta_1 c_M}{\eta_M c_1} \sin(\lambda_1). \end{cases}$$

As a result, the coefficients A_M^-, B_M^-, and similarly A_M^{*-} and B_M^{*-}, simplify to

$$(34)\qquad A_M^- = \cos(\lambda_1) \quad \text{and} \quad B_M^- = \frac{\eta_1 c_M}{\eta_M c_1} \sin(\lambda_1),$$

$$(35)\qquad A_M^{*-} = -\sin(\lambda_1) \quad \text{and} \quad B_M^{*-} = \frac{\eta_1 c_M}{\eta_M c_1} \cos(\lambda_1).$$

The parameters to be used in numerical experiments (see Figure 10 and Figure 14) are

$$d_0 = 0, \quad d_1 = l_1/2, \quad d_j = l_1/2 + \sum_{h=2}^{j} l_h, \quad b_0 = 0, \quad b_j = l_1/2, \quad j = 1, 2, \ldots, M,$$

that is,

$$\lambda_1^+ = \lambda_1^- = \lambda_1 = \frac{\omega l_1}{2c_1}, \quad \lambda_j^+ = \lambda_j = \frac{\omega l_j}{c_j}, \quad \lambda_j^- = 0, \quad j = 1, 2, \ldots, M.$$

Finally, by substituting (34) and (35) in the recursive dispersion relation (30) for the quasi-symmetric configuration, we obtain the recursive dispersion relation (36) for the general configuration of the unit cell made of M layers/materials.

$$(36)\quad \boxed{\begin{aligned} 2\cos(qa) = \tfrac{c_1 \eta_M}{c_M \eta_1}\big[&(A_M^+ \tfrac{\eta_1 c_M}{\eta_M c_1} \cos\lambda_1 + \cos\lambda_1 B_M^{*+} - A_M^{*+} \tfrac{\eta_1 c_M}{\eta_M c_1} \sin\lambda_1 + \sin\lambda_1 B_M^+) \cos(\lambda_M) \\ &+ (-A_M^+ \sin\lambda_1 - \cos\lambda_1 A_M^{*+} + B_M^+ \tfrac{\eta_1 c_M}{\eta_M c_1} \cos\lambda_1 - B_M^{*+} \tfrac{\eta_1 c_M}{\eta_M c_1} \sin\lambda_1) \sin(\lambda_M)\big]. \end{aligned}}$$

Here the coefficients with a (+) superscript are generated using the recursive relations (21)–(23) and (25)–(26). In summary, we established in (36) a recursive dispersion relation for the general configuration of the unit cell made of M layers/materials. This formula was obtained by considering the quasi-symmetric cell configuration as an essential intermediate step. For cell configurations that are already quasi-symmetric or symmetric, the recursive dispersion relation expressed by (30) is more suitable, as it involves fewer recursive steps.

5. Numerical results: Comparison of the central expansion approach with the method by Shen and Cao given in [18].

5.1. Three-material cell ($M = 3$). We consider a cell composed of three distinct materials: concrete, nickel alloy, and steel. The lengths of the layers are 0.2 m, 0.25 m, and 0.3 m, respectively. The general cell diagram in Figure 6 illustrates the unit cell used in [18] for $M = 3$. In the central expansion approach, a shifting and renumbering of the layers in the original cell takes place. The resulting cell is illustrated by the renumbered cell diagram given in Figure 6 with $N = \lceil \frac{M}{2} \rceil + 1 = \lceil \frac{3}{2} \rceil + 1 = 3$, hence the need for a fake interface. As a result, the cell to be used with the central expansion approach has five layers and the materials for each layer are steel, nickel, concrete, concrete, and steel. Notice the introduction of a fake interface on the original layer of concrete. The material parameters (elastic modulus and density) are given in the appendix.

The two graphs of the dispersion relation obtained using the central expansion approach given in (29), and Shen and Cao's formulas in [18], overlap in Figure 12, demonstrating the consistency between the two methods. The recursive relation generating the coefficients $A_N^{\pm}, B_N^{\pm}, A_N^{*\pm}, B_N^{*\pm}$ of (29) is programmed in Maple [15]. The dispersion relation graph is then obtained using the *implicitplot* Maple command. Determining more accurately the values of the circular frequency ω for a given value of q, including the band ends with $qa = 0$ or $qa = \pi$, is a difficult root-finding problem. As discussed in [11], [12], due to (4), it is only necessary to consider the wave number q limited to the domain $0 \leq q \leq \pi/a$. As seen in Figure 12, the dispersion relation graph displays a banded frequency spectrum using the reduced zone scheme for the wave number q, with the circular frequency ω in the ordinate. The banded frequency spectrum is composed of pass or propagating bands and stop bands. Over the interval $0 \leq q \leq \pi/a$, bands of permissible frequencies appear, separated by forbidden bands (creating bandgaps), at which frequencies no Floquet waves can be propagated. This band structure of pass and nonpass bands shows the dispersive properties of the medium. Similar comments can be made for the other dispersion relation graphs.

5.2. Four-material cell ($M = 4$). Similarly, we consider a cell composed of four distinct materials: steel, aluminum, concrete, and nickel alloy. The lengths of the layers are 0.15 m, 0.1 m, 0.4 m, and 0.2 m, respectively. As before, the general cell diagram in Figure 6 illustrates the unit cell used in [18], when $M = 4$. The cell used for the central expansion approach is illustrated by the renumbered cell diagram given in Figure 6 with $N = \lceil \frac{M}{2} \rceil + 1 = \lceil \frac{4}{2} \rceil + 1 = 3$. Notice that unlike the three-material case, in this case we do not need to add a fake interface because the number of layers after the shift is already odd. The materials for each of the five layers are nickel alloy, steel, aluminum, concrete, and nickel alloy. We plot the dispersion relation using both methods, as shown in Figure 13. Notice again how well the two graphs overlap.

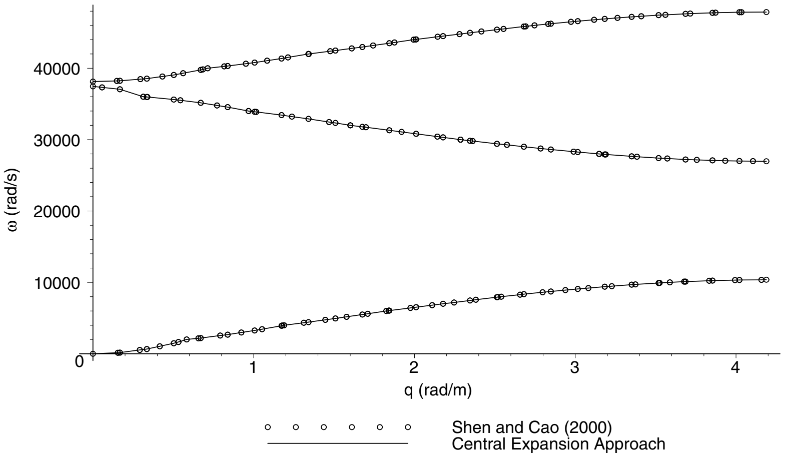


Fig. 12. *Dispersion graphs for a three-material cell: Central expansion approach versus Shen and Cao* [18], *using the reduced zone scheme for the wave number q.*

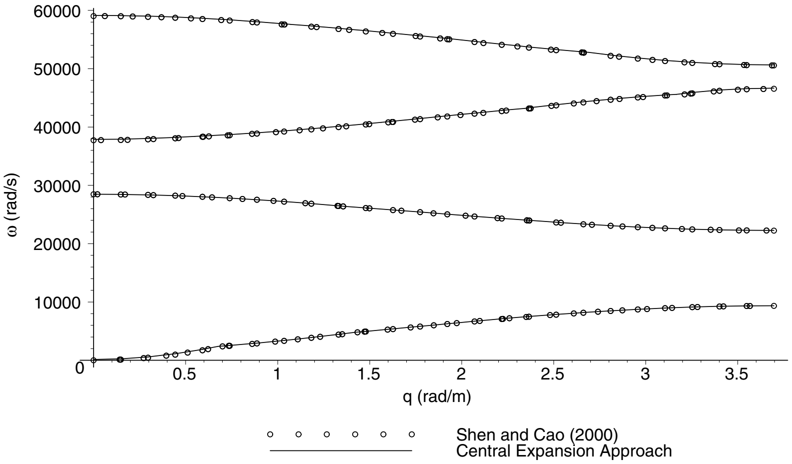


Fig. 13. *Dispersion graphs for a four-material cell: Central expansion approach versus Shen and Cao* [18], *using the reduced zone scheme for the wave number q.*

6. Numerical results: Quasi-symmetric limiting approach.

6.1. Comparison of the quasi-symmetric limiting approach with the method by Shen and Cao in [18]. Here we consider a unit cell composed of five layers arranged in a general configuration, shown in Figure 14. We choose aluminum as material 1, nickel alloy as material 2, concrete as materials 3 and 5, and steel as material 4, with the following layer thicknesses: $l_1 = 0.1$ m, $l_2 = 0.05$ m, $l_3 = 0.4$ m, $l_4 = 0.2$ m, and $l_5 = 0.25$ m. The material parameters (elastic modulus and density) can be found in the appendix. The two graphs displaying the dispersion relation using the quasi-symmetric limiting approach given in (36), and Shen and Cao's formulas in [18], overlap in Figure 15.

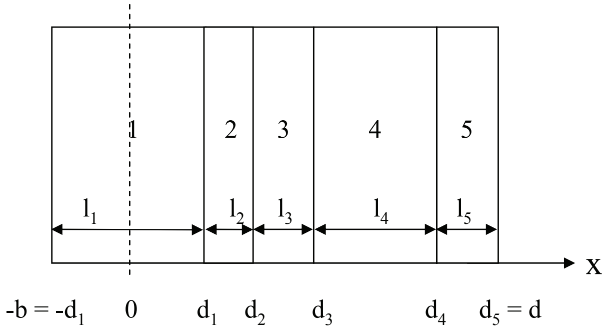


Fig. 14. *General configuration of a five-layer cell.*

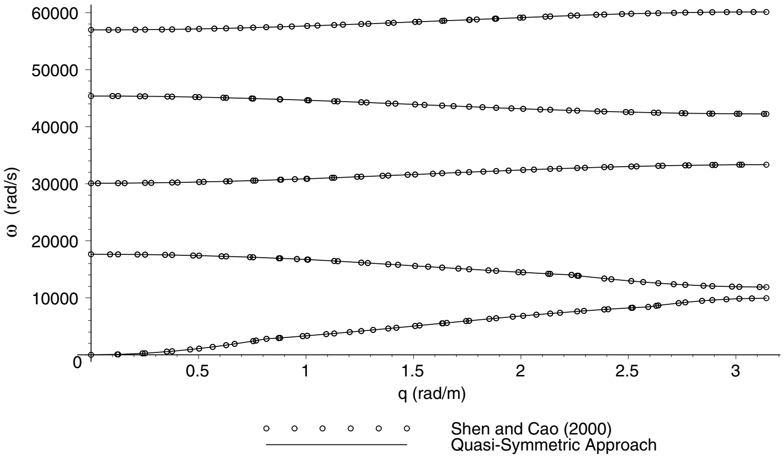


Fig. 15. *Dispersion graphs for a five-layer cell: Quasi-symmetric limiting approach versus Shen and Cao* [18], *using the reduced zone scheme for the wave number q.*

6.2. Comparison of both of our approaches with experimental results: A simplified one-dimensional model. The dispersion relation and sound attenuation through three-dimensional structures composed of periodically arranged cubic cells were experimentally measured in [13]. The cubic single cell consisted of a 1 cm diameter spherical core made of lead, coated with a 2.5 mm layer of silicone rubber. The coated spheres were periodically arranged in a $8 \times 8 \times 8$ cubic crystal with lattice constant of 1.55 cm, and with epoxy as the surrounding matrix material. The cross section of the cell is displayed on the upper part of Figure 16. The one-dimensional, three-material symmetric cell model, shown in the lower part of Figure 16, may be viewed, as suggested by Wang et al. [23], as a simplified one-dimensional counterpart of the three-dimensional structure studied in [13]. In our one-dimensional model, lead is considered as material 1, silicone rubber as material 2, and epoxy as material 3 with the following layer lengths: 1 cm (central), with 0.25 cm and 0.025 cm on each side. The material parameters (elastic modulus and density) are included in the appendix. Our intent here is to illustrate the fact that our dispersion relations correctly predict bandgaps due to a so-called localized resonance phenomenon that has been observed in three-dimensional ternary systems [13]; the localized resonance

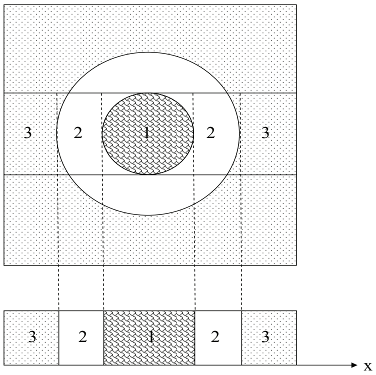


FIG. 16. *One-dimensional unit cell motivated from the three-dimensional cubic structure (cross section) studied in* [13].

can reduce the magnitude of the bandgap by two orders of magnitude relative to that caused by Bragg scattering, and this observation has spurred renewed interest by the acoustic bandgap community for design of acoustic attenuators. In [6], Bragg's law $f = n * v/2a$, for example, predicts a scattering frequency of about 62.5 KHz using steel scatterers embedded in an epoxy matrix with a lattice constant of 20 mm. Here, v is the longitudinal wave speed of the matrix, and a is the distance between the centers of the scatterers ($n = 1, 2, 3, \ldots$).

Liu et al. [13] measured acoustic transmission T as a function of frequency (250 Hz to 1600 Hz) by placing a receiving transducer at the center of their sonic crystals with an external sound source. The lowest values of T correspond to wave frequencies that are attenuated by the structure, whereas the highest values of T correspond to wave frequencies that easily propagate throughout the structure. Their experiments reveal that peak transmission frequencies are located at $f = 600$ Hz and $f = 1600$ Hz. Between these two frequencies, the transmission coefficient is low. The few measurements made for low frequency ($f \simeq 300$ Hz) suggest that waves with low frequencies easily propagate through the structure.

Due to the symmetric cell configuration of the one-dimensional simplified model given in Figure 16, the central expansion approach and the quasi-symmetric limiting approach share the same dispersion relation formula (29). The graph in Figure 17 displays the dispersion relation predicted by (29) and obtained using the *implicitplot* Maple command. Here we use frequency f on the ordinate instead of the circular frequency ω. Determining more accurately the values of the frequency f for a given value of q, including the band ends with $qa = 0$ or $qa = \pi$, is a difficult root-finding problem. As seen in Figure 17, between the frequency range of 0–2000 Hz, the graph exhibits a pass band for low frequencies under 140 Hz, and what appear to be four additional narrow pass bands centered approximately at $f = 873$, $f = 950$, $f = 1353$, and $f = 1907$ Hz. Closer inspection of an expanded view of the second pass band shows that it is centered at $f = 873.5$ Hz (Figure 18). Expanded views of the third and fourth bands (not shown here) located at approximately 950 Hz and 1353 Hz are essentially flat to within numerical roundoff, hence the group velocity at these frequencies is zero, i.e., $v_g = \frac{df}{dq} = 0$.

In conclusion, our one-dimensional model appears to qualitatively predict the acoustic response of the three-dimensional ternary structure, with the two narrow pass

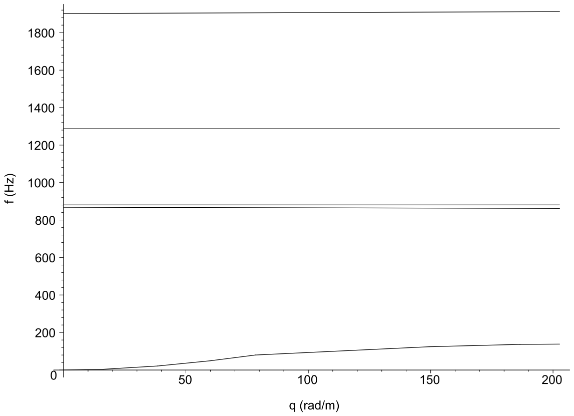


Fig. 17. *Graph of the dispersion relation for one-dimensional symmetrical cell model motivated from the three-dimensional model of* [13]*, using the reduced zone scheme for the wave number q.*

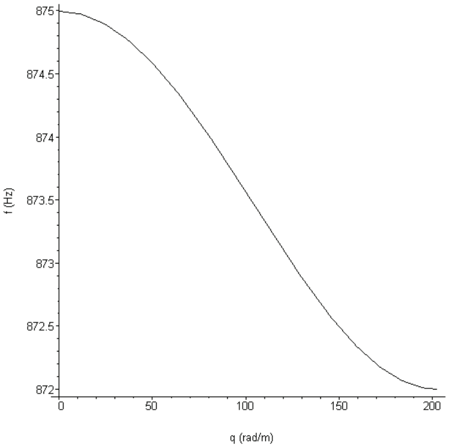


Fig. 18. *Expanded view of the second pass band illustrated in Figure* 17.

bands corresponding to the peaks and the two large bandgaps corresponding to the ranges of frequencies that are highly attenuated. Our numerical simulations predict bandgaps with a lattice constant two orders of magnitude smaller than the relevant wavelength, suggesting a so-called localized resonance phenomenon already observed in three-dimensional ternary systems [13]. However, the locations of the bandgaps do not match those that are experimentally observed. This is to be expected, considering the fact that our band structure equations are derived for infinite one-dimensional periodic elastic media whereas the experiments in [13] were conducted on a finite three-dimensional structure, with bandgaps only partially developed.

Appendix. The following are the elastic modulus η and density ρ of the materials selected for the numerical experiments:

1. Concrete: $\eta = 33 \cdot 10^9$ Pa and $\rho = 2400$ kg/m^3.
2. Steel: $\eta = 210 \cdot 10^9$ Pa and $\rho = 7800$ kg/m^3.
3. Aluminum: $\eta = 69 \cdot 10^9$ Pa and $\rho = 2710$ kg/m^3.

4. Nickel Alloy: $\eta = 214 \cdot 10^9$ Pa and $\rho = 8130$ kg/m^3.
5. Lead: $\eta = 40.8 \cdot 10^9$ Pa and $\rho = 11600$ kg/m^3.
6. Silicone rubber: $\eta = 117500$ Pa and $\rho = 1300$ kg/m^3.
7. Epoxy: $\eta = 4.4 \cdot 10^9$ Pa and $\rho = 1180$ kg/m^3.

Acknowledgment. The authors would like to acknowledge the two reviewers of this paper for their valuable suggestions and constructive comments.

REFERENCES

[1] A. Ben-Menahem and S. J. Singh, *Seismic Waves and Sources*, Springer-Verlag, New York, 1981.

[2] L. Brillouin, *Wave Propagation in Periodic Structures*, Dover, New York, 1946.

[3] W. Cao and W. Qi, *Plane wave propagation in finite* 2-2 *composites*, J. Appl. Phys., 78 (1995), pp. 4627–4632.

[4] P. Goupillaud, *An approach to inverse filtering of near-surface layer effects from seismic records*, Geophys., 26 (1961), pp. 754–760.

[5] N. A. Haskell, *The dispersion of surface waves on multilayered media*, Bull. Seismol. Soc. Amer., 43 (1953), pp. 17–34.

[6] M. Hirsekorn, P. P. Delsanto, N. K. Batra, and P. Matic, *Modelling and simulation of acoustic wave propagation in locally resonant sonic materials*, Ultrasonics, 42 (2004), pp. 231–235.

[7] M. I. Hussein, G. M. Hulbert, and R. A. Scott, *Dispersive elastodynamics of* 1*D banded materials and structures: Analysis*, J. Sound Vibration, 289 (2006), pp. 779–806.

[8] M. I. Hussein, G. M. Hulbert, R. A. Scott, and K. Saitou, *Multiobjective evolutionary optimization of periodic layered materials for desired wave dispersion characteristics*, Struct. Multidiscip. Optim., 31 (2006), pp. 60–75.

[9] E. R. Kanasewich, *Time Sequence Analysis in Geophysics*, University of Alberta Press, Edmonton, Alberta, 1973.

[10] M. S. Kushwaha, *Classical band structure of periodic elastic composites*, Internat. J. Modern Phys. B, 10 (1996), pp. 977–1094.

[11] E. H. Lee, *A survey of variational methods for elastic wave propagation analysis in composites with periodic structures*, in Dynamics of Composite Materials, E. H. Lee, ed., ASME, New York, 1972, pp. 122–138.

[12] E. H. Lee and W. H. Yang, *On waves in composite materials with periodic structure*, SIAM J. Appl. Math., 25 (1973), pp. 492–499.

[13] Z. Liu, X. Zhang, Y. Mao, Y. Y. Zhu, Z. Yang, C. T. Chan, and P. Sheng, *Locally resonant sonic materials*, Science, 289 (2000), pp. 1734–1736.

[14] W. Magnus and S. Winkler, *Hill's Equation*, John Wiley, New York, 1996.

[15] *Maple* 12 *User Manual*, Waterloo Maple, Waterloo, ON, Canada, 2008.

[16] W. L. Pilant, *Elastic Waves in the Earth*, Elsevier Science, New York, 1979.

[17] Lord Rayleigh, *On the maintenance of vibrations of forces of double frequency, and on the propagation of waves through a medium endowed with a periodic structure*, Phil. Mag., 24 (1887), pp. 145–159.

[18] M. Shen and W. Cao, *Acoustic bandgap formation in a periodic structure with multilayer unit cells*, J. Phys. D, 33 (2000), pp. 1150–1154.

[19] J. W. C. Sherwood and A. W. Trorey, *Minimum-phase and related properties of the response of horizontally stratified absorptive earth to plane acoustic waves*, Geophys., 30 (1965), pp. 191–197.

[20] C. T. Sun, J. D. Achenbach, and G. Herrmann, *Continuum theory for a laminated medium*, J. Appl. Mech., 35 (1968), pp. 467–475.

[21] W. T. Thomson, *Transmission of elastic waves through a stratified solid medium*, J. Appl. Phys., 21 (1950), pp. 89–93.

[22] S. Treitel and E. A. Robinson, *Seismic wave propagation in layered media in terms of communication theory*, Geophys., 31 (1966), pp. 17–32.

[23] G. Wang, D. Yu, J. Wen, Y. Liu, and X. Wen, *One-dimensional phononic crystals with locally resonant structures*, Phys. Lett. A, 327 (2004), pp. 512–521.

[24] S. Yang, J. H. Page, Z. Liu, M. L. Cowan, C. T. Chan, and P. Sheng, *Ultrasound tunneling through* 3*D phononic crystals*, Phys. Rev. Lett., 88 (2002), article 104301.

SIAM J. APPL. MATH.
Vol. 69, No. 3, pp. 690–703

THE SUPPRESSION OF FOUR-WAVE MIXING BY RANDOM DISPERSION[*]

RUDY L. HORNE[†], CHRISTOPHER K. R. T. JONES[‡], AND TOBIAS SCHÄFER[§]

Abstract. Pairwise interactions of optical pulses lead to the creation of four-wave mixing (FWM) products in the presence of periodic damping and amplification. In this manuscript, we examine how these FWM products grow in the presence of small to moderate random dispersion. Namely, we show that (i) the growth of the FWM products in the presence of white noise is inversely proportional to the noise strength $\bar{D}$, confirmed by both analytical and numerical results; (ii) the FWM products as a function of $\bar{D}$ obey a gamma-type probability distribution function; and (iii) the presence of either white or Ornstein–Uhlenbeck (OU) noise has a similar influence on the growth of these FWM products. This work shows that small random dispersion can effectively mitigate the deleterious effects of FWM in wavelength-division–multiplexed optical communications systems for either sech-type or Gaussian-type input pulses.

Key words. four-wave mixing, random dispersion, resonance condition, noise strength

AMS subject classifications. 35Q51, 41A60

DOI. 10.1137/070680539

1. Introduction.

In optical fiber transmission systems, there has been a great push to increase the throughput in both single and multiple channel fiber lines. One technique employed is wavelength-division multiplexing (WDM). This technique allows for the propagation of multiple pulses on different channels along a single fiber line. Problems arise when pulses from one channel interact with pulses from a different channel. Indeed, these pulse interactions create by-products that interfere with other pulses within the fiber line. In this manuscript, we examine these by-products from pairwise collisions of optical pulses in different channels under the influence of random system noise.

One of the most common scenarios of pulse interactions is that of pairwise interactions. In this case, two pulses from different frequency channels come together and interact (i.e., collide) with one another by virtue of their different speeds. In linear systems, this type of interaction is described by a superposition of waves. In nonlinear systems (i.e., WDM optical fiber systems), however, these pairwise interactions result in the creation of by-products (known as four-wave mixing (FWM) products) that can interfere with the propagation of other pulses in the fiber system. These products result from a resonance that is excited due to the nonlinearity of the index of refraction, the so-called Kerr nonlinearity. In this paper, we study what happens when random noise is added to the system and how this affects the growth of these by-products.

[*] Received by the editors January 22, 2007; accepted for publication (in revised form) July 1, 2008; published electronically December 17, 2008. This research was partially supported by the NSF under grants DMS-0205496 and FRG-0073923.

http://www.siam.org/journals/siap/69-3/68053.html

[†] Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510 (horne@math.fsu.edu). This author's research was partially supported by the Carolina Postdoctoral Program for Faculty Diversity through the Office of the Vice Chancellor for Research and Graduate Studies at the University of North Carolina at Chapel Hill and by the FYAP program at Florida State University.

[‡] Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3250 (ckrtj@email.unc.edu).

[§] Department of Mathematics, College of Staten Island, Staten Island, NY 10314 (tobias@math.csi.cuny.edu). This author's research was partially supported by the grant PSCREG-37-454 of the CUNY Research Foundation.

The model equation that describes pulse propagation in optical fibers is the nonlinear Schrödinger (NLS) equation with periodic damping and amplification. In this work, we use the NLS equation to derive a linear partial differential equation which models the evolution of the FWM growth. This model is then used to determine an analytical expression for the energy (L^2-norm) of the FWM growth in the presence of random dispersion. This result is compared with calculation of the energy of the FWM products via simulations of the NLS equation. These simulations are done in the presence of white noise as well as for Ornstein–Uhlenbeck (OU) noise. We then numerically show the probability distribution function associated with the energy of the FWM growth.

2. Motivation and paper outline. Pairwise nonlinear interactions between pulses in different frequency channels give rise to several effects. Two major effects are (i) collision-induced timing jitter, which results in permanent frequency shifts between colliding pulses (see [20, 3, 13]), and (ii) FWM product generation and growth that results in a permanent waveform that may interact with other optical pulses (see [5, 19, 11, 6]).

Fiber impurities, mostly due to fiber construction and implementation, can be viewed as small random perturbations to the fiber. In the past, these perturbations have often been considered a negative effect on the optical system in that they lead to pulse distortions or even pulse destruction: solitons [2] and dispersion-managed solitons [1, 17, 21] broaden under the influence of random perturbations until they disintegrate. Fortunately, these effects take place over very large distances, and that is why it is possible to use these solitons as bit carriers in terrestrial communication networks. In this paper, we adopt an entirely different point of view by showing that, with respect to other phenomena, namely FWM growth and deterministic resonances, random variations of the dispersion profile can have a beneficial impact concerning pulse interactions. The central issue we address is how random dispersion affects FWM growth. In previous work [12], we showed that FWM products are reduced in the presence of random dispersion. The numerical results presented in [12] showed that weak random dispersion reduces FWM product growth. This reduction can be compared with other methods such as dispersion management (see [9, 16, 15, 7]). We also showed that the resonance condition associated with FWM growth is affected by the presence of random dispersion. In this manuscript, we continue our investigation of FWM product production in the presence of random dispersion. We show that the energy associated with the FWM products (i.e., the L^2-norm associated with the FWM products) gives rise to an integral which one can analyze asymptotically. This asymptotic treatment allows us to give a quantitative comparison between the analytical treatment of the FWM growth and Monte Carlo simulations of the full NLS equation with random dispersion.

The outline for this manuscript is as follows. In the first two sections, we outline the derivation of a linear PDE which describes the FWM growth associated with pairwise collisions of optical fibers. This linear PDE is further reduced to a linear ODE which is used throughout the analysis. Next, we use the linear ODE model to derive an expression for the L^2-norm associated with FWM growth in the presence of random dispersion. This analysis shows that the energy of the FWM products depends inversely on the noise strength, $\bar{D}$. This result is verified via simulations of the full NLS equation for these FWM products in the third section. We consider both white and OU noise. In the fourth section, we give a numerical description of the probability distribution function associated with the energies for these FWM products. Finally, we summarize these results and state some conclusions.

3. FWM growth model.

3.1. Derivation of the FWM growth model: The NLS equation. We begin this analysis with the NLS equation, which includes varying dispersion, damping, and amplification:

$$iE_z + \frac{\beta(Z)}{2} E_{TT} + \gamma |E|^2 E = -i\alpha E. \tag{3.1}$$

$E(Z,T)$ is the envelope of the electric field, where Z and T are the propagation and time variables measured in km and ps, respectively; $\beta(Z)$ is the variable dispersion profile; and the loss and damping/amplification coefficients are given by γ and α.

Let us introduce the dimensionless variables $\zeta = Z/z_*$, $\tau = T/t_*$, $|\beta_{\text{av}}|$, and $Q(\zeta,\tau) = E(Z,T)/\sqrt{P_*}$, where z_*, t_*, $|\beta_{\text{av}}|$, and P_* denote the characteristic length scale, characteristic time scale, the average dispersion value, and the peak power, respectively. We define length scales associated with the dispersion ($z_{\beta_{\text{av}}}$) and the nonlinearity (z_{NL}) given by $z_{\beta_{\text{av}}} = t_*/|\beta_{\text{av}}|$ and $z_{NL} = 1/\gamma P_*$. We also define the dimensionless amplifier spacing, $z_a = L_a/z_*$, where L_a is the physical amplifier spacing taken to be 40km. Under this change of variables, (3.1) becomes

$$iQ_\zeta + \frac{z_* d(\zeta)}{2 z_{\beta_{\text{av}}}} Q_{\tau\tau} + \frac{z_*}{z_{NL}} |Q|^2 Q = -i\Gamma Q, \tag{3.2}$$

where $\Gamma = z_*\alpha$ and $d(Z) = \beta(Z) z_{\beta_{\text{av}}}/{t_*}^2$. A more convenient form of (3.2) is found with a further change of variables and matching certain parameters.

Setting $Q(\zeta,\tau) = \sqrt{g(\zeta)}\, u(\zeta,\tau)$ and taking $z_* = z_{\beta_{\text{av}}} = z_{NL}$, one determines the following form of the NLS equation:

$$iu_z + \frac{d(z)}{2} u_{tt} + g(z)|u|^2 u = 0. \tag{3.3}$$

Here, we have relabeled our variables $(\zeta,\tau) \longrightarrow (z,t)$, $g(z)$ is an exponential function with period z_a (nonlinear coefficient), and $d(z)$ is the dispersion profile of the form $d(z) = d_{\text{av}} + F(z)$, where, d_{av} is the average dispersion and $F(z)$ is either some prescribed function or a white noise process. Throughout this paper, we work with (3.3) and take $F(z)$ to be a white noise process. If one sets $g(z) = 1$ and $F(z) = 0$, (3.3) supports soliton solutions of the form

$$u(z,t) = A\text{sech}[A(t - \Omega d_{\text{av}} z + T)] \exp[i(A^2 - \Omega^2)z/2] \exp(i\Omega t). \tag{3.4}$$

Here, A is the pulse amplitude, $\Omega = \pi c t_* \Delta\lambda/\lambda^2$ is the frequency offset, c is the speed of light, t_* is the characteristic time scale, λ is the wavelength of the soliton, $\Delta\lambda$ is the channel spacing, and T is the timing offset for the soliton pulse. We also take typical parameter values as follows: $L_a = 40\text{km}$, $z_* = 400\text{km}$, $\lambda = 1550\text{nm}$, $\alpha = 0.025\text{km}^{-1}$, and $P_* = 2\text{mW}$, which implies that $z_a = L_a/z_* = 0.1$ and $\Gamma = 10.0$. Also, the characteristic time $t_* = \tau_{pulse}/1.763$, where the number 1.763 denotes the full-width half maximum of the ideal soliton and τ_{pulse}= 28ps. We note that $\Omega = 3.9$ corresponds to a channel spacing of $\Delta\lambda = 0.62\text{nm}$.

3.2. Derivation of the FWM growth model: Model equation. We briefly outline the derivation of the model PDE that describes the growth of the FWM products. The presentation given in this section is a brief summary of work done in [11].

We assume that our pulses can be written as a sum of the two interacting pulses and the FWM products [4, 11]:

$$u(z,t) \simeq u_1(z,t) + u_2(z,t) + u_{\text{fwm}}(z,t). \tag{3.5}$$

Upon substitution of (3.5) into (3.3), one derives the following linear PDE:

$$iq_z + \frac{d(z)}{2} q_{tt} = -g(z) {u_2}^2 {u_1}^*. \tag{3.6}$$

Here, we have that $q(z,t) \equiv u_{221}(z,t)$ is the FWM product growth associated with frequency 3Ω, $u_j(z,t)$ are the input pulses, $d(z)$ is the dispersion profile, which is some prescribed function, and $g(z)$ is the periodic damping and amplification.

We note that on the right-hand side (RHS) of (3.6) the forcing term has a rapidly varying piece which we now factor out in the form

$$q(z,t) = Q(z,t) \exp\left(3i\Omega t - i\frac{\Omega^2}{2}\int_0^z \, d(z')dz'\right). \tag{3.7}$$

Plugging this into our linear PDE for $q(z,t)$, one derives the following equation for $Q(z,t)$:

$$iQ_z + \frac{d(z)}{2}(Q_{tt} + 6i\Omega Q_t - 2(2\Omega)^2 Q) = -g(z) {u_{20}}^2 {u_{10}}^*, \tag{3.8}$$

where

$$ {u_{20}}^2 {u_{10}}^* \equiv {u_2}^2 {u_1}^* \exp\left(-3i\Omega t + i\frac{\Omega^2}{2}\int_0^z \, d(z')dz'\right) \tag{3.9}$$

and $g(z)$ is as described in (3.3). We note that for the case of initially well-separated pulses, $\Delta\Omega = 2\Omega \gg 1$, where 2Ω is the difference in the frequency offset for the two pulses. For this case, one notes that in (3.8) we have $(2\Omega)^2|Q| \gg |Q_t|, |Q_{tt}|$. Here, we have assumed that our input pulses are well separated (i.e., $(2\Omega) \gg 1$). Typical values for this case occur when $\Omega \geq 2$ (see [20]), which is the case of interest throughout this manuscript. This allows us to further reduced our model FWM equation to the form

$$iQ_z - d(z)(2\Omega)^2 Q = -g(z) u_{20}^2 u_{10}^*. \tag{3.10}$$

In this model, the input pulses on the right-hand side of (3.10) can be either classical solitons or Gaussian-type pulses. Equation (3.10) is the FWM growth model used in the analysis throughout this manuscript. In the next section, we incorporate noise into the dispersion coefficient, $d(z)$ in (3.10), and analyze the FWM growth in this case via its L^2-norm.

4. FWM growth and random dispersion.

4.1. Random dispersion in an optical fiber. In this paper, we take the dispersion, $d(z)$, to have the following form:

$$d(z) = d_{av} + \xi(z), \tag{4.1}$$

where $\xi(z)$ is a white noise process with

$$\langle \xi(z) \rangle = 0 \qquad \text{and} \qquad \langle \xi(z)\xi(z') \rangle = \bar{D}\ \delta(z - z'). \tag{4.2}$$

The noise strength, $\bar{D}$, is taken to be small (i.e., $0 < \bar{D} \ll 1$) for our problem of interest. We consider small to moderate noise strength so as to not destroy the input pulses. Since the input soliton pulses decay slowly on the length scale, $z_{degr} = 1/\bar{D}$ (see [8]), we work with distances on the order of $z \ll z_{degr}$. Also, we take the RHS of (3.8) as independent of the randomness given by (4.1)–(4.2). We assume that, to leading order, the input pulses are roughly fixed relative to the input noise. We now present an analysis for the growth of FWM products as a function of the noise strength $\bar{D}$ and the frequency offset Ω.

4.2. FWM growth: Analytical results.

In this section, we derive an equation which captures the behavior of the FWM products in the presence of a white noise process. These results will be shown to be in quantitative agreement with what is seen numerically.

We restate the reduced model equation for FWM growth:

$$iQ_z - d(z)(2\Omega)^2 Q = -g(z){u_{20}}^2{u_{10}}^*. \tag{4.3}$$

Here, $u_{j0}(z,t)$ denotes the form of the input pulse for j = 1,2. Our input pulses can have one of the following forms:

- Sech-type input pulse:

$$u_{j0}(z,t) = A_j \operatorname{sech}[A_j(t - \Omega_j d_{av} z - T_j)] \exp[i({A_j}^2 - {\Omega_j}^2)z/2] e^{i\Omega_j t}, \tag{4.4}$$

  where $\Omega_2 = -\Omega_1 \equiv \Omega$ and $T_1 = -T_2 \equiv T_0$. We take $A_1 = A_2 \equiv A$ unless otherwise specified in this manuscript.

- Gaussian-type input pulse:

$$u_{j0}(z,t) = \frac{\alpha_j}{\sqrt{2\pi\beta_j}} \exp[-(t - \Omega_j d_{av} z - T_j)^2/2\beta_j] \exp[i({\alpha_j}^2 - {\Omega_j}^2)z/2] e^{i\Omega_j t}, \tag{4.5}$$

  where $\Omega_2 = -\Omega_1 \equiv \Omega$ and $T_1{=}-T_2 \equiv T_0$. Similar to the case for sech-type pulses, we take $\alpha_1{=}\alpha_2 \equiv \alpha$ and $\beta_1{=}\beta_2 \equiv \beta$ unless otherwise specified.

We note that, strictly speaking, the input pulses given by (4.4)–(4.5) should incorporate randomness via the dispersion within the argument of the input pulses. For the distances discussed in this manuscript ($z \ll z_{degr}$), the pulses are roughly fixed relative to the random dispersion.

Also, for either type of input pulses, the pairwise collision will occur at the point

$$z_{\text{collision}} \equiv z_{coll} = \frac{T_1 - T_2}{d_{av}(\Omega_2 - \Omega_1)} = \frac{T_0}{d_{av}\Omega} \neq 0. \tag{4.6}$$

The collision point is a function of the timing and frequency offsets, and all results are derived with this in mind. We also note that this analysis applies to complete collisions (i.e., initially well-separated input pulses). If the input pulses are not well separated, then the collision is an incomplete collision, which leads to much larger inelastic effects (see [18, 13]).

Going back to (4.3) and solving for $Q(z,t)$ yields

$$Q(z,t) = ie^{-i(2\Omega)^2 \int_0^z d(z')dz'} \int_0^z dz'\ g(z'){u_{20}}^2{u_{10}}^*(z',t)e^{i(2\Omega)^2 \int_0^{z'} d(s)ds}, \tag{4.7}$$

where $\int_0^z d(z')dz' = d_{av}z + W(z)$ with $W(z) \equiv \int_0^z \xi(s)ds$. We recall that $g(z)$ is a periodic function in z and therefore is expressible as a Fourier series:

$$g(z) = \sum_{m=-\infty}^{\infty} g_m \, e^{-i2\pi m z/z_a}, \qquad g_m = \frac{\Gamma z_a}{\Gamma z_a - im\pi}, \tag{4.8}$$

where z_a is the dimensionless amplifier spacing. Plugging (4.8) into (4.7), one finds

$$Q(z,t) = ie^{-i(2\Omega)^2 \int_0^z \, d(z')dz'} \sum_{m=-\infty}^{\infty} g_m \, I_m(z,t), \tag{4.9}$$

where the integral within the sum is given as

$$I_m(z,t) \equiv \int_0^z dz' \, e^{i\psi_m z'} e^{i(2\Omega)^2 W(z')} [u_{20}^2 u_{10}^*](z',t) = \int_0^z dz' \, e^{i\chi_m(z')} [u_{20}^2 u_{10}^*](z',t). \tag{4.10}$$

Here, $\psi_m \equiv \bar{\lambda}^2/2 + (2\Omega)^2 \, d_{av} - 2\pi m/z_a$ is a resonance condition which determines where the largest FWM products are produced as a function of Ω for $m = 1, 2, 3, \ldots$ (see [11, 4]), $\bar{\lambda}$ is a parameter (usually the amplitude) which depends on the form of the input pulses, and $W(z') = \int_0^{z'} \xi(s')ds'$. Here, $W(z)$ is a normally distributed random variable with

$$\langle W(z) \rangle = 0 \qquad \text{and} \qquad \langle W^2(z) \rangle = \bar{D}z. \tag{4.11}$$

Upon examination of the argument of the $\exp[i\chi_m(z')]$ term in the integrand of $I_m(z,t)$, it is clear that the resonance condition is modified in the presence of the random variable $W(z)$:

$$\chi_m(z') \equiv \psi_m z' + (2\Omega)^2 W(z') = \left(\frac{\lambda^2}{2} + (2\Omega)^2 d_{\mathrm{av}} - \frac{2\pi m}{z_a} \right) z' + (2\Omega)^2 W(z'). \tag{4.12}$$

Equation (4.12) implies that the resonance condition is affected by the random variable $W(z')$. If one views $\chi_m(z')$ as a random function, then one can determine that $\langle \chi_m(z') \rangle = \psi_m z'$ since $\langle W(z') \rangle = 0$. In this work, we present a general analysis of how the size of the FWM products is affected by randomness. We see this through examination of the quantity $\langle |Q(z,t)|^2 \rangle$.

Now, going back to (4.9)–(4.10), one finds

$$|Q(z,t)|^2 = \sum_{m,n=-\infty}^{\infty} g_m \, g_n^* \, I_m(z,t) I_n{}^*(z,t). \tag{4.13}$$

At this point, we can determine the L^2-norm of $Q(z,t)$ (i.e., $\int_{-\infty}^{\infty} |Q(z,t)|^2 dt$). This is done by interchanging the integral and summand as follows:

$$\int_{-\infty}^{\infty} |Q(z,t)|^2 dt = \sum_{m,n=-\infty}^{\infty} g_m \, g_n^* \int_{-\infty}^{\infty} I_m(z,t) I_n{}^*(z,t) dt. \tag{4.14}$$

As concerns the growth of the FWM products, we are interested in the quantity $\langle \int_{-\infty}^{\infty} |Q(z,t)|^2 dt \rangle$. For our problem, the randomness comes in only through the z-dependence, so t can be treated as a parameter relative to the averaging over z. We have that

$$\left\langle \int_{-\infty}^{\infty} |Q(z,t)|^2 dt \right\rangle = \int_{-\infty}^{\infty} \langle |Q(z,t)|^2 \rangle dt, \tag{4.15}$$

which tells us that the quantity of interest is $\langle |Q(z,t)|^2\rangle$. Using (4.14), we have

$$\langle |Q(z,t)|^2\rangle = \sum_{m,n=-\infty}^{\infty} g_m\ g_n^*\ \langle I_m(z,t) I_n{}^*(z,t)\rangle. \tag{4.16}$$

Again, $I_m(z,t)$ has the form

$$I_m(z,t) \equiv \int_0^z dz'\ f(z',t)\ e^{i\psi_m z'}\ e^{i(2\Omega)^2 W(z')}, \tag{4.17}$$

where

$$f(z,t) \equiv [u_{20}^2 u_{10}^*](z,t)\ e^{-i(A^2-\Omega^2)z/2} \quad \text{and} \quad \psi_m \equiv (2\Omega)^2 d_{av} + \frac{\lambda^2}{2} - \frac{2\pi m}{z_a}. \tag{4.18}$$

By examination of the product of the integrals $I_n(z,t)\ I_m{}^*(z,t)$ and using (4.17), we find

$$I_m(z,t)\ I_n{}^*(z,t) = \int_0^z \int_0^z f(s,t) f(s',t)\ e^{i\psi_m s} e^{-i\psi_n s'} e^{i(2\Omega)^2[W(s)-W(s')]}\ ds\ ds'. \tag{4.19}$$

Here, the region of integration corresponds to integrating over the square box: $0 \le s, s' \le z$. Defining the variable $\tilde{W} \equiv W(s) - W(s')$, it is well known that $\tilde{W}$ is also a normally distributed random variable with mean and variance given by (see [10])

$$\langle \tilde{W}\rangle = 0 \qquad \text{and} \qquad \langle \tilde{W}^2\rangle = \bar{D}\ |s-s'|. \tag{4.20}$$

Using elementary calculus and the fact that $\langle e^{k\tilde{W}}\rangle = e^{\frac{k^2\langle \tilde{W}^2\rangle}{2}}$ for $\tilde{W}$ normally distributed with zero mean and k as some constant value, one finds

$$\begin{aligned} \langle I_m(z,t)\ I_n{}^*(z,t)\rangle &= \int_0^z \int_0^z ds' ds\ f(s,t) f(s',t)\ e^{i\psi_m s}\ e^{-i\psi_n s'}\ e^{-(2\Omega)^4 \bar{D}|s-s'|} \\ &= \sum_{j=1}^{4} \int\int_{R_j} ds' ds\ f(s,t) f(s',t)\ e^{i\psi_m s}\ e^{-i\psi_n s'}\ e^{-(2\Omega)^4 \bar{D}|s-s'|}. \end{aligned} \tag{4.21}$$

Here, R_j denotes the following regions of integration: (i) R_1: $0 \le s, s' \le z_{coll}$, (ii) R_2: $0 \le s \le z_{coll}$ and $z_{coll} \le s' \le z$, (iii) R_3: $z_{coll} \le s, s' \le z$, and (iv) R_4: $z_{coll} \le s \le z$ and $0 \le s' \le z_{coll}$. Again, z_{coll} is the collision point of the two pulses (either sech-type or Gaussian-type) stated in (4.6).

We can evaluate each of the integrals in (4.21) with the use of the following transformation:

$$u = s' - s, \quad v = s + s' \quad \Longrightarrow \quad s' = \frac{u+v}{2}, \quad s = \frac{v-u}{2}. \tag{4.22}$$

Applying (4.22) to the regions R_j for $j = 1, 2, 3, 4$, we convert (4.21) into

$$\langle I_m(z,t)\ {I_n}^*(z,t)\rangle$$

$$= \int_0^{z_{coll}} du\, G_1(u)\ \cos\left[\frac{(\psi_m+\psi_n)u}{2}\right] e^{-xu} + \frac{1}{2}\int_0^{z-z_{coll}} du\, F_1(u) e^{-i(\psi_m+\psi_n)u/2} e^{-xu}$$

$$+ \frac{1}{2}\int_0^{z_{coll}} du\, F_1(u) e^{i(\psi_m+\psi_n)u/2} e^{-xu} + \int_0^{z-z_{coll}} du\, G_2(u)\cos\left[\frac{(\psi_m+\psi_n)u}{2}\right] e^{-xu}$$

$$+ \frac{1}{2}\int_{z_{coll}}^{z} du\, F_2(u) e^{-i(\psi_m+\psi_n)u/2} e^{-xu} + \frac{1}{2}\int_{z-z_{coll}}^{z} du\, F_2(u) e^{i(\psi_m+\psi_n)u/2} e^{-xu}$$

$$+ \frac{1}{2}\int_{z-z_{coll}}^{z_{coll}} du\, K_1(u) e^{-i(\psi_m+\psi_n)u/2} e^{-xu} + \frac{1}{2}\int_{z_{coll}}^{z-z_{coll}} du\, K_2(u) e^{i(\psi_m+\psi_n)u/2} e^{-xu}, \tag{4.23}$$

where the parameter $x \equiv (2\Omega)^4\ \bar{D}$, $z \geq z_{coll}$, and the quantities $G_1(u)$, $G_2(u)$, $F_1(u)$, $F_2(u)$, $K_1(u)$, and $K_2(u)$ are defined below:

$$G_1(u) \equiv \int_u^{2z_{coll}-u} dv\ H(u,v), \qquad G_2(u) \equiv \int_{2z_{coll}+u}^{2z-u} dv\ H(u,v),$$

$$F_1(u) \equiv \int_{2z_{coll}-u}^{2z_{coll}+u} dv\ H(u,v), \qquad F_2(u) \equiv \int_u^{2z-u} dv\ H(u,v),$$

$$K_1(u) \equiv \int_{2z_{coll}-u}^{2z-u} dv\ H(u,v), \qquad K_2(u) \equiv \int_u^{2z_{coll}+u} dv\ H(u,v). \tag{4.24}$$

The quantity $H(u,v)$ incorporates information about the two pulses:

$$H(u,v) \equiv f\left(\frac{v+u}{2},\ \ t\right)\ f\left(\frac{v-u}{2},\ \ t\right)\ e^{i(\psi_m-\psi_n)v/2}. \tag{4.25}$$

Typical values for our problem of interest are $\Omega \simeq 3.9$ and $0.005 \leq \bar{D} \leq 0.03$ (corresponding to 0.5% to 3% of variation with respect to the average dispersion set to $d_{av} = 1.0$), which corresponds to a large x value (i.e., $x \gg 1$). One can apply Laplace's method to examine each of the integrals in (4.23) for large x. If one takes $z \longrightarrow \infty$, the resulting asymptotic integral is given by

$$\lim_{z\longrightarrow\infty}\ \langle I_m(z,t)\ {I_n}^*(z,t)\rangle$$

$$\simeq \frac{1}{(2\Omega)^4\bar{D}}\ \left[\int_0^{\infty} dv\ H(u=0,\ v)\ +\ \frac{1}{2}\ \int_{z_{coll}}^{3z_{coll}} dv\ H(u=z_{coll},\ v)\right]. \tag{4.26}$$

Again, we note that this analysis applies to either sech-type or Gaussian-type pulses. One can bound each of the integrals in (4.26) and then show that the quantity $\langle I_m(z,t){I_n}^*(z,t)\rangle$ decays like $1/(2\Omega)^4\bar{D}$, where Ω is the frequency offset and $\bar{D}$ represents the noise strength arising from the random dispersion given in (4.1) and (4.2). In the next section, we compare (4.26) with numerical integration of (3.3), which is the full NLS equation with periodic damping and amplification, to show quantitative agreement between the two approaches.

4.3. FWM growth: Numerical results for white noise process. In the previous section, we showed analytically that the dominant contribution to the L^2-norm for FWM growth in the presence of a random dispersive term decayed like $1/(2\Omega)^4\bar{D}$,

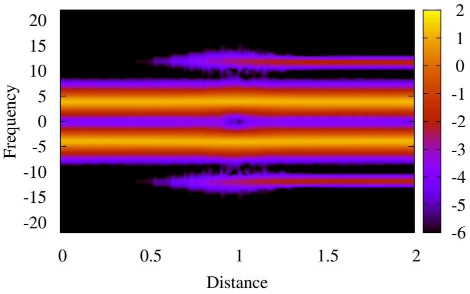


Fig. 1. *Contour plot of the spectrum of a signal for a pairwise interaction of sech-type pulses in the absence of noise in the dispersion coefficient: The dispersion here is simply a constant function (i.e., $d(z) = 1.0$) to show the presence of FWM growth side-bands. The color corresponds to the logarithm of the amplitude.*

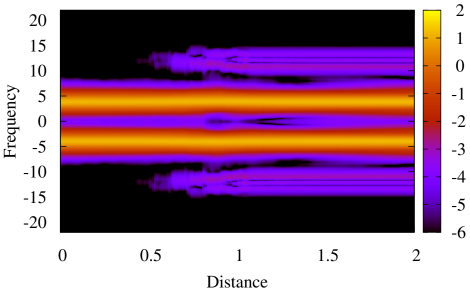


Fig. 2. *Contour plot of the spectrum of a signal for a pairwise interaction of sech-type pulses in the presence of white noise. The noise strength is given by $\bar{D} = 0.03$. The color corresponds to the logarithm of the amplitude, indicating that the side-bands are smoothed out and overall much weaker than in the system without noise.*

where Ω and $\bar{D}$ represent the frequency offset and noise strength, respectively. Again, the FWM growth results from a pairwise collision of either sech- or Gaussian-type pulses (see (4.4)–(4.5)) and is best observed in the frequency domain.

We chose to examine the pairwise interaction of sech-type pulses. The reason for this choice was simply one of convenience for the numerics. All of the results presented here can be applied to Gaussian-type pulses without loss of the basic features concerning FWM products. In Figure 1, we show the frequency spectrum for the propagating signal, which consists of two interacting pulses that are well separated in the frequency domain, in the absence of dispersive noise. Due to the interaction, they develop FWM that occurs as side-bands in the frequency domain. The simulation clearly shows how these bands are first created and then persist. This simulation was obtained by solving (3.3) using a standard second-order split-step method.

Running the same simulation as shown in Figure 1 but allowing for small randomness in the dispersion coefficient $d(z)$ (see (4.1)–(4.2)), considerable suppression of the FWM products arises. In Figure 2, the noise strength is chosen to be $\bar{D} = 0.03$

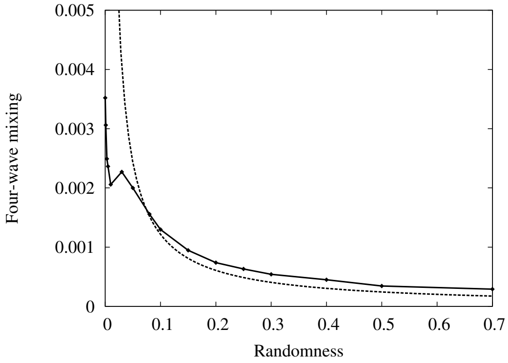


Fig. 3. *This is a plot of* $R = ||u_{FWM}||^2/||u_{sol}||^2$ *versus* $\bar{D}$*, where* R *is the ratio of the FWM energy* ($||u_{FWM}||^2$) *to one of the input sech-type pulse's energy* ($||u_{sol}||^2$) *and* $\bar{D}$ *is the noise strength parameter. The solid line is obtained by integration of the full NLS equation for* 1000 *Monte Carlo simulations at each* $\bar{D}$ *value, taking* $z = 2.5$ *dimensionless units. The dashed line represents the result obtained using* (4.15), (4.16)*, and* (4.26).

and shows that the amplitude of the side-bands is much lower. This indicates that the four-wave products are reduced in a noisy system. In our simulations, we use a standard split-step algorithm in order to integrate the full NLS equation (3.3). At each step in the evolution variable z, we randomly draw a number that approximates the white noise in the dispersion term. In order to model the white noise effectively, this random number is drawn from a normal distribution with the variance $\sigma^2 = \bar{D}/dz$ at each evolution step dz.

In section 4.2 we derived an expression for the L^2-norm for the FWM products given in (4.15), (4.16), and (4.26). In Figure 3, we plot the L^2-norm of the FWM products (relative to the L^2-norm of one of the sech-type pulses) versus the noise strength $\bar{D}$, given the soliton parameters $\Omega = 3.9$ and $T_0 = 4.0$. The plot is generated by solving the full NLS equation for propagation of the pulses to some finite distance along the fiber after the collision ($z = 2.5$ dimensionless units, where the collision occurs at $z_{coll} = 1.027$). A certain distance after the collision of the pulses (at $z = 2.5$ dimensionless units), we measure the value $R = ||u_{FWM}||^2/||u_{sol}||^2$ for a fixed noise strength $\bar{D}$. Figure 3 tells us the value of R for a range of $\bar{D}$ values, averaging the Monte Carlo simulations for 1000 realizations per run in z per $\bar{D}$ value. We compare the numerical results obtained by integration of the full NLS equation with the sum that arises from solving the linear PDE model for the FWM growth (see (4.16)). This figure shows that the FWM growth decays like $1/(2\Omega)^4\bar{D}$, which is shown in both numerical simulations as well as from the analytical results.

The sum in (4.16) was evaluated using twenty terms, with the asymptotic formula given by (4.26). From (4.23)–(4.26), it is clear that the dominant terms arise when the indices m and n are equal. Since we are dealing with sech-type (or Gaussian-type) pulses, the contributions to (4.16) can be shown numerically to be small as m and n become large. Practically speaking, twenty terms give sufficient accuracy (up to at least five digits of numerical accuracy).

In this section, we have shown both numerically and analytically that the FWM products decay inversely with the noise strength parameter $\bar{D}$. This confirms the initial evidence of this phenomenon seen in [12]. Another question to be asked is What is the probability distribution function of the R value (ratio of L^2-norms) associated

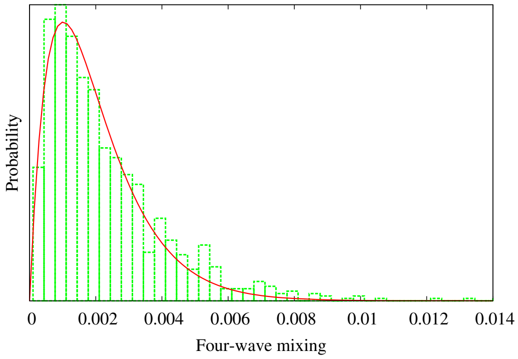


FIG. 4. *This is a plot of the probability distribution function versus* $R = ||u_{FWM}||^2/||u_{sol}||^2$ *given* $\bar{D} = 0.03$ *and* $\Omega = 3.9$.

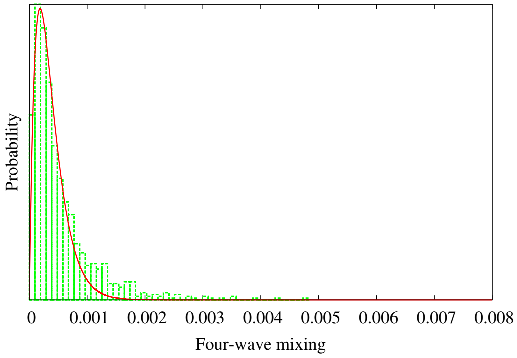


FIG. 5. *This is a plot of the probability distribution function versus* $R = ||u_{FWM}||^2/||u_{sol}||^2$ *given* $\bar{D} = 0.30$ *and* $\Omega = 3.9$.

with these FWM products? We show some numerical results and propose a model for the probability distribution function.

4.4. FWM growth and its probability distribution function. In the previous section, we found that qualitatively the FWM products decrease like $1/(2\Omega)^4\bar{D}$ (see (4.26)). Indeed, we can numerically determine the probability distribution function of the FWM products for a set value of the noise strength $\bar{D}$. Once this is done, we can determine a best fit curve for the probability distribution function associated with the ratio of L^2-norms for the FWM products.

We begin by solving the NLS equation (3.3), where $g(z)$ is a periodic function, $d(z)$ is given by (4.1)–(4.2), and sech-type input pulses are used as given in (4.4). We numerically integrate the NLS equation for 1000 Monte Carlo simulations, setting the appropriate constants Ω, T_0, and $\bar{D}$. Once the histogram is numerically calculated, we fit this curve to a gamma-type probability distribution function.

In Figures 4 and 5, we plot histograms of the probability distribution functions as functions of R for different $\bar{D}$ values. These figures are obtained upon numerical integration of the full NLS equation (3.3) for $\Omega \simeq 3.9$, $\bar{D} = 0.03$, and $\bar{D} = 0.30$ for 1000 Monte Carlo simulations. The results in these figures confirm what was predicted in Figure 3; as we increase the value of the noise strength $\bar{D}$, there is a significant

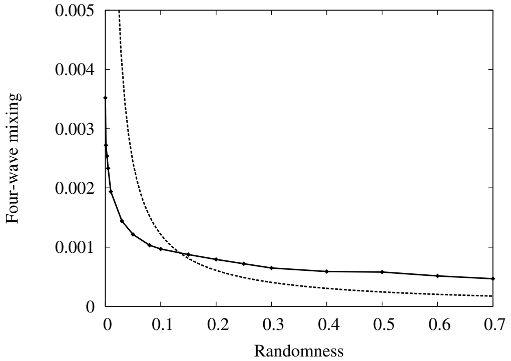


FIG. 6. *This is a plot of $R = ||u_{FWM}||/||u_{sol}||$ versus the noise strength parameter $\bar{D}$ in the presence of OU noise in the dispersive term. The solid line is obtained by integration of the full NLS equation for* 1000 *Monte Carlo simulations for a fixed $\bar{D}$ value for $z = 2.5$ dimensionless units. The dashed line is the evaluation of the sum for the white noise case using* (4.15), (4.16), *and* (4.26).

reduction in the FWM products. These probability distribution functions also give us an indication of how the values of R are distributed for a given Ω and $\bar{D}$. From Figures 4 and 5, it appears that the results in these figures are best approximated by a Gamma-type distribution. By fitting the data using a least-squares approach, one can show that the probability distribution function is approximated by

$$f(x) = \alpha x e^{-\beta x}, \tag{4.27}$$

where $\alpha \approx 9.2356 \times 10^5$ and $\beta \approx 986.315$ when $\bar{D} = 0.03$, and $\alpha \approx 2.42866 \times 10^7$ and $\beta \approx -5366.57$ for 0.30, respectively. Here, x represents the binning of the R values, and β, n, and m depend on the noise strength $\bar{D}$.

4.5. FWM: Numerical results for a correlated noise process. In section 4.3, we showed that the FWM growth is inversely proportional to the noise strength $\bar{D}$ in the presence of a white noise process in the dispersion coefficient. Indeed, the results displayed in Figure 3 summarized both the numerical and analytical results derived for the case of white noise in the dispersive term. A question that we raise is Is the FWM growth affected in a similar manner for different types of noise processes? We briefly give some numerical results for the case of OU noise. In this case, $\xi(z)$ in (4.1) is created by solving the corresponding stochastic equation:

$$d\xi(z) = -k\xi(z)dz + \sqrt{\bar{D}_{OU}}dW, \tag{4.28}$$

where $\bar{D}_{OU}$ is the appropriately scaled diffusion constant for the OU process and k^{-1} is the correlation length (see [10]).

Figure 6 is a plot of the L^2-norm of the FWM products, relative to the L^2-norm of one of the input pulses, versus the noise strength, $\bar{D}$. This plot was generated in a similar manner to that of Figure 3. We solve the full NLS equation for propagation of the pulses to some finite distance along the fiber after the collision ($z = 2.5$ dimensionless units). After this, we measure the value of $R = ||u_{FWM}||^2/||u_{sol}||^2$ (the ratio of the L^2-norm of the FWM product and one of the input pulses) at $z = 2.5$ for a fixed noise strength $\bar{D}$. Again, we determine the values of R for a range of $\bar{D}$ values, averaging the Monte Carlo simulations for 1000 simulations per run in z per $\bar{D}$ value. We plot the analytical result determined for the white noise process to show that, qualitatively, the FWM products decay inversely to the noise strength $\bar{D}$. For the case of short-correlated noise in the form of an OU process, the effect of the noise on the FWM differs slightly from the white noise case (again, see Figure 3), but we again see an efficient suppression of the FWM product.

5. Summary of results and conclusions. In this manuscript, we have examined how randomness in the dispersion term affects the growth of FWM products. We have shown that in the presence of a white-noise process, the decay of the L^2-norm for the FWM products is proportional to $1/(2\Omega)^4\bar{D}$, where Ω is the frequency offset of the input pulse and $\bar{D}$ is the noise strength parameter. We have confirmed this result in two ways: (i) by the development of an analytical model that describes the FWM growth and (ii) using Monte Carlo simulations on the full NLS equation. We have also given a qualitative description for the probability distribution function associated with the FWM products for a fixed noise strength. Furthermore, we have shown that, qualitatively, these results are the same when the white-noise process is replaced by an Ornstein-Uhlenbeck process.

Acknowledgments. R. L. Horne would like to acknowledge useful discussions with R. Camassa and R. McLaughlin at the University of North Carolina at Chapel Hill. Also, the authors are grateful for the comments by the reviewer of this paper; the paper is a much stronger paper for this reason.

REFERENCES

[1] F. Kh. Abdullaev and B. B. Baizakov, *Disintegration of a solition in a dispertion-managed optical communication line with random parameters*, Opt. Lett., 25 (2000), p. 93.

[2] F. Kh. Abdullaev, J. C. Bronski, and G. Papanicolaou, *Soliton perturbations and the random Kepler problem*, Phys. D, 135 (2000), pp. 369–386.

[3] M. J. Ablowitz, G. Biondini, S. Chakravarty, and R. L. Horne, *On timing jitter in wavelength-division multiplexed soliton systems*, Opt. Comm., 150 (1998), pp. 305–318.

[4] M. J. Ablowitz, G. Biondini, S. Chakravarty, and R. L. Horne, *Four-wave mixing in dispersion-management return-to-zero systems*, J. Opt. Soc. Amer. B Opt. Phys., 20 (2003), pp. 831–845.

[5] M. J. Ablowitz, G. Biondini, S. Chakravarty, R. B. Jenkins, and J. R. Sauer, *Far-wave mixing in wavelength-division-multiplex soliton systems: Damping and amplification*, Opt. Lett., 21 (1996), p. 1646.

[6] M. J. Ablowitz, G. Biondini, S. Chakravarty, R. B. Jenkins, and J. R. Sauer, *Four-wave mixing in dispersion-management return-to-zero systems: Ideal fibers*, J. Opt. Soc. Amer. B Opt. Phys., 14 (1997), pp. 1788–1794.

[7] N. S. Bergano, C. R. Davidson, M. A. Mills, P. C. Corbett, R. Menges, J. L. Zyskind, J. W. Shulhoff, A. K. Srivastava, and C. Wolf, in Optical Amplifiers and Their Applications, A. Willner, M. Zervas, and S. Sasaki, eds., Vol. 16 of OSA Trends in Optics and Photonics Series, Optical Society of America, Washington, DC, 1997, postdeadline paper PD-9.

[8] M. Chertkov, Y. Chung, A. Dyachenko, I. Gabitov, I. Kolokolov, and V. Lebedev, *Shedding and interaction of solitons in a weakly disordered optical fiber*, Phys. Rev. E, 67 (2003), paper 036615-1.

[9] F. Favre, D. LeGuen, and T. Georges, *Experimental evidence & pseudo-periodical soliton propogation in dispersion-managed link,* Electron. Lett., 34 (1998), pp. 1868–1869.

[10] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer Ser. Synergetics, Springer, New York, 2004.

[11] R. L. Horne, *Collision-Induced Timing Jitter and Four-Wave Mixing in Wavelength-Division Multiplexing Soliton Systems*, Ph.D. thesis, Applied Mathematics Department, University of Colorado at Boulder, Boulder, CO, 2001.

[12] R. L. Horne, C. K. R. T. Jones, and T. Schäfer, *The effects of weak randomness on pulse interactions and four-wave mixing products,* Phys. D, 205 (2005), pp. 70–79.

[13] D. J. Kaup, B. A. Malomed, and J. Yang, *Collision-induced pulse timing jitter in a wavelength-division-multiplexing system with strong dispersion management*, J. Opt. Soc. Amer. B Opt. Phys., 16 (1999), pp. 1628–1635.

[14] R. J. Larsen and M. L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, Prentice–Hall, Englewood Cliffs, New Jersey, 1986.

[15] F. Liu, J. Bennike, S. Dey, C. Rasmussen, H. Mikkelsen, P. Mamyshev, D. Gapontse, and V. Ivshin, in Optical Fiber Communication Conference (OFC), Vol. 70 of OSA Trends in Optics and Photonics Series, Optical Society of America, Washington, DC, 2002.

[16] L. F. Mollenauer, P. V. Mamyshev, J. Gripp, M. J. Neubelt, N. Mamysheva, L. Grüner-Nielson, and T. Veng, *Demonstration of massive wavelength-division multiplexing over transoceanic distances by use of dispersion-managed solitons*, Opt. Lett., 25 (2000), pp. 704–706.

[17] B. A. Malomed and A. Berntson, *Propagation of an optical pulse in a fiberlink with random-dispersion management*, J. Opt. Soc. Amer. B Opt. Phys., 18 (2001), pp. 1243–1251.

[18] B. Malomed, G. D. Peng, and P. L. Chu, *Soliton wavelength-division multiplexing system with channel-isolating notch fitters*, Opt. Lett., 24 (1999), p. 1100.

[19] P. V. Mamyshev and L. F. Mollenauer, *Pseudo-phase-matched four-wave mixing soliton wavelength-division multiplexing transmission*, Opt. Lett., 21 (1996), p. 396.

[20] L. F. Mollenauer, S. G. Evangelides, and J. P. Gordon, J. Lightwave Tech., *Wavelength division multiplexing with solitons in ultra-long distance transmission using lumped amplifiers*, 9 (1991), pp. 362–367.

[21] T. Schäfer, R. O. Moore, and C. K. R. T. Jones, *Pulse propagation in media with deterministic and random dispersion variations*, Opt. Comm., 214 (2002), pp. 353–362.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 704–719

SPECTRUM OF A LINEARIZED AMPLITUDE EQUATION FOR ALTERNANS IN A CARDIAC FIBER*

SHU DAI† AND DAVID G. SCHAEFFER†

Abstract. Under rapid periodic pacing, cardiac cells typically undergo a period-doubling bifurcation in which action potentials of short and long duration alternate with one another. If these action potentials propagate in a fiber, the short-long alternation may suffer reversals of phase at various points along the fiber, a phenomenon called (spatially) discordant alternans. Either stationary or moving patterns are possible. Using a weak approximation, Echebarria and Karma proposed an equation to describe the spatiotemporal dynamics of small-amplitude alternans in a class of simple cardiac models, and they showed that an instability in this equation predicts the spontaneous formation of discordant alternans. To study the bifurcation, they computed the spectrum of the relevant linearized operator numerically, supplemented with partial analytical results. In the present paper we calculate this spectrum with purely analytical methods in two cases where a small parameter may be exploited: (i) small dispersion or (ii) a long fiber. From this analysis we estimate the parameter ranges in which the phase reversals of discordant alternans are stationary or moving.

Key words. spectrum, amplitude equation, cardiac alternans

AMS subject classifications. 35B32, 92C30

DOI. 10.1137/070711384

1. Introduction.

Alternans, a period-doubling bifurcation of action potential durations in rapidly paced cardiac cells, has been implicated as a precursor of ventricular fibrillation [6, 5, 10, 2]. When such action potentials propagate in tissue, their short-long alternation may suffer reversals of phase; such *discordant* alternans pose even higher arrythmogenic risks. Since ventricular fibrillation accounts for 1/6 of all deaths in the USA [1, 7], great importance attaches to understanding these phenomena.

Echebarria and Karma [3] proposed a weakly nonlinear description of the one-dimensional evolution of discordant alternans in cardiac models[1] for which each action potential duration (APD) is a function of only the previous diastolic interval (DI). To set the context, suppose a cardiac fiber of length L is stimulated periodically at its $x = 0$ end, say with period B (mnemonic for basic cycle length, which has the acronym BCL). It is assumed that each stimulus successfully generates an action potential that propagates down the fiber. Let $A_k(x)$ be the duration of the kth action potential at the position x along the fiber. For slow stimulation, say $B > B_{\text{crit}}$, the propagating action potentials become identical after a transient: i.e., $\lim_{k\to\infty} A_k(x)$ exists and is independent of x. In studying pacing with $B < B_{\text{crit}}$, Echebarria and

*Received by the editors December 19, 2007; accepted for publication (in revised form) August 7, 2008; published electronically December 17, 2008. This work was supported by the National Institutes of Health under grant 1R01-HL-72831 and by the National Science Foundation under grant PHY-0549259. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siap/69-3/71138.html

†Department of Mathematics and Center for Nonlinear and Complex Systems, Duke University, Durham, NC 27708 (sdai@math.duke.edu, dgs@math.duke.edu).

[1] No cardiology background is required to read this paper if (1.2) is accepted as a given. An appendix, written primarily for mathematicians, reviews the context in which (1.2) arises.

Karma make the ansatz

$$A_k(x) = A_{\text{crit}} - \delta A + (-1)^k a(x,t), \tag{1.1}$$

where A_{crit} is the APD when pacing with period $B = B_{\text{crit}}$, δA is the average shortening of APD resulting from decreasing B below B_{crit}, and $a(x,t)$ is the amplitude of alternans, assumed slowly varying. Because of this slow-variation assumption, one may study the evolution of a with respect to a continuous time t that interpolates between the times $t = kB$, $k = 0, 1, 2, \ldots$, when stimuli are applied. Nondimensionalizing the time by B_{crit}, they derive the evolution equation for $a(x,t)$:

$$\partial_t a = \sigma a + \xi^2 \partial_{xx} a - w \partial_x a - \frac{1}{\Lambda} \int_0^x a(x',t)dx' - ga^3, \tag{1.2}$$

where σ is the bifurcation parameter, which is dimensionless and proportional to $B_{\text{crit}} - B$; Λ, w, and ξ are all positive parameters in units of length, which are derived from the equations of the cardiac model; and the nonlinear term $-ga^3$ limits growth after the onset of linear instability. Boundary conditions

$$\partial_x a(0,t) = 0, \quad \partial_x a(L,t) = 0 \tag{1.3}$$

are imposed on (1.2).

Of course, $a \equiv 0$ is a solution of (1.2)–(1.3), but it loses stability as σ increases. Bifurcation analysis of this equation requires knowing the eigenvalues Ω_n of the linear operator that maps a function $a(x)$ to

$$\xi^2 \partial_{xx} a - w \partial_x a - \frac{1}{\Lambda} \int_0^x a(x',t)dx', \tag{1.4}$$

subject to Neumann boundary conditions. All of these eigenvalues lie in the (stable) left-half plane. The eigenvalue(s) with the largest real part, say $\Omega_{\max}$, determines the character of the solution of (1.2) at the onset of bifurcation—a stationary pattern if $\Omega_{\max}$ is real, a moving pattern if it is complex.

In this paper we extend the results of [3] by calculating the spectrum of (1.4) with purely analytical means in two limiting cases: small dispersion and a long fiber. In particular, it follows from our analysis that in a long fiber $\Omega_{\max}$ is real if, modulo terms that are $\mathcal{O}(L^{-2})$,

$$\Lambda^{-1} \le C \frac{w^3}{\xi^4}, \tag{1.5}$$

where

$$C = \frac{1}{64}\left(71 + 17\sqrt{17}\right) \approx 2.205, \tag{1.6}$$

and $\Omega_{\max}$ is complex otherwise.

2. The eigenvalue problem.

Let us begin by nondimensionalizing (1.4). The parameters ξ, w, and Λ, like L, all have the units of length. Thus we define a set of new parameters

$$\bar{w} = w/\xi, \qquad \bar{\Lambda} = \Lambda/\xi, \qquad \bar{L} = L/\xi, \tag{2.1}$$

and (1.4) can be written as

$$\frac{d^2a}{d\bar{x}^2} - \bar{w}\frac{da}{d\bar{x}} - \bar{\Lambda}^{-1}\int_0^{\bar{x}} a(\bar{x}')d\bar{x}', \tag{2.2}$$

where $\bar{x} = x/\xi$. For further scaling, we define

$$\bar{\bar{x}} = \bar{w}\cdot\bar{x}, \qquad \bar{\bar{L}} = \bar{w}\cdot\bar{L}, \qquad \bar{\bar{\Lambda}} = \bar{\Lambda}\cdot\bar{w}^3 \tag{2.3}$$

and an operator

$$\mathscr{L} = \frac{d^2a}{d\bar{\bar{x}}^2} - \frac{da}{d\bar{\bar{x}}} - \bar{\bar{\Lambda}}^{-1}\int_0^{\bar{\bar{x}}} a(\bar{\bar{x}}')d\bar{\bar{x}}'. \tag{2.4}$$

One observes that (1.4) equals $\bar{w}^2\cdot\mathscr{L}a$.

The analysis in sections 2–4 below uses dimensionless variables, but we nonetheless *shall omit all the bars* in (2.4). Suppose that $a(x)$ is an eigenfunction of (2.4) with eigenvalue Ω: i.e.,

$$\mathscr{L}a = \Omega\, a, \quad \text{with} \quad \mathrm{a}'(0) = 0, \quad \mathrm{a}'(\mathrm{L}) = 0. \tag{2.5}$$

To eliminate the integral term in (2.4), we differentiate this equation (but not the boundary conditions (B.C.)) with respect to x to obtain

$$\begin{cases} a''' - a'' - \Lambda^{-1}a = \Omega\, a', \\ a'(0) = 0, \\ a'(L) = 0, \\ a''(0) = \Omega\, a(0). \end{cases} \tag{2.6}$$

The additional B.C. comes from evaluating the eigenvalue equation, before differentiation, at $x = 0$. A function of the form $a(x) = e^{\kappa x}$ satisfies the ODE in (2.6) if

$$\kappa^3 - \kappa^2 - \Omega\kappa - \Lambda^{-1} = 0. \tag{2.7}$$

If $\kappa_1, \kappa_2, \kappa_3$ are the roots of (2.7), then this equation may be reformulated as

$$1 = \kappa_1 + \kappa_2 + \kappa_3, \tag{2.8}$$

$$\Omega = -(\kappa_1\kappa_2 + \kappa_2\kappa_3 + \kappa_3\kappa_1), \tag{2.9}$$

$$\Lambda^{-1} = \kappa_1\kappa_2\kappa_3. \tag{2.10}$$

Assuming that the roots κ_1, κ_2, κ_3 are distinct, we seek a solution of (2.6) of the form $a(x) = \sum_1^3 C_i e^{\kappa_i x}$. The three B.C.s in (2.6) give a homogeneous linear system for the unknown coefficients C_i. For this system to possess a nontrivial solution, we need

$$\det\begin{pmatrix} \kappa_1 & \kappa_2 & \kappa_3 \\ \kappa_1 e^{\kappa_1 L} & \kappa_2 e^{\kappa_2 L} & \kappa_3 e^{\kappa_3 L} \\ \Omega - \kappa_1^2 & \Omega - \kappa_2^2 & \Omega - \kappa_3^2 \end{pmatrix} = 0. \tag{2.11}$$

Thus $\Omega \in \mathbb{C}$ is an eigenvalue of $\mathscr{L}$ if there exist a triple κ_1, κ_2, κ_3, no two of them equal, such that the four equations (2.8)$-$(2.11) are satisfied.

If $\Lambda^{-1} > 0$ so that each root κ_i is nonzero, (2.11) may be reformulated as follows. By (2.7), $\Omega - \kappa_i^2 = -\kappa_i - \Lambda^{-1}\kappa_i^{-1}$. Substituting this expression into the third row of (2.11) and manipulating the determinant, we obtain

$$\det\begin{pmatrix} 1 & 1 & 1 \\ e^{\kappa_1 L} & e^{\kappa_2 L} & e^{\kappa_3 L} \\ \kappa_1^{-2} & \kappa_2^{-2} & \kappa_3^{-2} \end{pmatrix} = 0. \tag{2.12}$$

Let us rule out possible multiple roots. First regarding a triple root, which by (2.8) must be $\kappa_1 = \kappa_2 = \kappa_3 = \frac{1}{3}$: In this case the general solution to (2.6) is of the form $a(x) = (C_0 + C_1 x + C_2 x^2)\, e^{x/3}$, and the only possible eigenvalue is $\Omega = -\frac{1}{3}$ by (2.9). Substituting $a(x)$ and Ω into (2.6), we find that there is no nontrivial solution and hence no eigenvalue. Now we assume $\kappa_1 = \kappa_2 \neq \kappa_3$ in (2.8)−(2.10), and therefore the general solution to (2.6) is $a(x) = C_1 e^{\kappa_1 x} + C_2 x e^{\kappa_1 x} + C_3 e^{\kappa_3 x}$. Inserting $a(x)$ into the boundary conditions in (2.6) and considering (2.8)−(2.10) with $\kappa_2 = \kappa_1$, we find that the existence of a nontrivial solution requires κ_1 to satisfy both of the following equations,

$$\kappa_1 L(4\kappa_1^2 - 3\kappa_1 - 1) + 2(1 - 2\kappa_1)^2 \cdot \left[e^{(1-3\kappa_1)L} - 1\right] = 0, \tag{2.13}$$

$$-2\kappa_1^3 + \kappa_1^2 = \Lambda^{-1} > 0, \tag{2.14}$$

and the possible eigenvalue is then given by

$$\Omega = 3\kappa_1^2 - 2\kappa_1. \tag{2.15}$$

Note that (2.13) only has isolated complex roots since its left-hand side is a holomorphic function. Thus (2.13)−(2.15) can provide only isolated eigenvalues in the Λ^{-1}-Ω plane, and a perturbation of the parameter Λ^{-1} will lead to the case when all κ_j's are different. So the case $\kappa_1 = \kappa_2 \neq \kappa_3$ can be obtained as limit of the case of distinct roots. In the analysis below we will see that the roots remain separated for Λ^{-1} small or L large.

3. Small dispersion:[2] $\Lambda^{-1} \ll 1$.

If $\Lambda^{-1} = 0$, then $\mathscr{L}$ has eigenvalues

$$\Omega_0^{(0)} = 0, \tag{3.1}$$

with the solution of (2.8)−(2.10) given by $\kappa_j = 0,\ 0,\ 1$, and

$$\Omega_n^{(0)} = -\frac{1}{4} - \left(\frac{\pi n}{L}\right)^2, \qquad n = 1, 2, \ldots, \tag{3.2}$$

with $\kappa_j = 0,\ \frac{1}{2} \pm \frac{in\pi}{L}$. We seek the first term in an expansion of Ω_n in powers of Λ^{-1}.

3.1. Perturbation of Ω_n, $n \geq 1$.

We prove that for Λ^{-1} small,

$$\Omega_n(\Lambda) = \Omega_n^{(0)} + \Omega_n^{(1)}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}), \tag{3.3}$$

where $\Omega_n^{(0)}$ is given by (3.2) and

$$\Omega_n^{(1)} = -\frac{2}{1 + 4\pi^2 n^2 L^{-2}}. \tag{3.4}$$

[2] We remind the reader that in this section and the next, Λ and L, without bars, refer to the dimensionless parameters defined by (2.3).

By the implicit function theorem, we may expand the solution of (2.8)–(2.10) and (2.12) as

$$\left\{\begin{aligned} &\kappa_{1,2} = \frac{1}{2} \pm \frac{in\pi}{L} + b_{1,2}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}),\\ &\kappa_3 = 0 + b_3\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}),\\ &\Omega_n = \Omega_n^{(0)} + \Omega_n^{(1)}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}). \end{aligned}\right. \tag{3.5}$$

Substituting into (2.8)−(2.10), we find from the $\mathcal{O}(\Lambda^{-1})$ terms

$$\left\{\begin{aligned} &b_1 + b_2 + b_3 = 0,\\ &\Omega_n^{(1)} = -\frac{b_1+b_2}{2} + \frac{in\pi}{L}(b_1 - b_2) - b_3,\\ &b_3 = \left(\frac{1}{4} + \frac{n^2\pi^2}{L^2}\right)^{-1}. \end{aligned}\right. \tag{3.6}$$

Substituting into (2.12), we find that the leading order term $\mathcal{O}(\Lambda^2)$ vanishes identically; from the next order term $\mathcal{O}(\Lambda)$ we deduce that $b_1 = b_2$. The relation (3.4) follows from this equation and (3.6).

3.2. Perturbation of $\boldsymbol{\Omega_0}$. We prove that for Λ^{-1} small,

$$\Omega_0(\Lambda) = \Omega_0^{(1)}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}), \tag{3.7}$$

where

$$\Omega_0^{(1)} = -\left(1 - \frac{L}{\exp(L) - 1}\right). \tag{3.8}$$

The expression analogous to (3.5) is complicated by the fact that, for $\Lambda^{-1} = 0$, $\kappa_1 = \kappa_2 = 0$ is a double root of (2.8)–(2.10). Thus we seek a Puisseux expansion

$$\left\{\begin{aligned} &\kappa_{1,2} = a_{1,2}\Lambda^{-1/2} + b_{1,2}\Lambda^{-1} + \mathcal{O}(\Lambda^{-3/2}),\\ &\kappa_3 = 1 + b_3\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}),\\ &\Omega_0 = \Omega_0^{(1)}\Lambda^{-1} + \mathcal{O}(\Lambda^{-2}). \end{aligned}\right. \tag{3.9}$$

First we substitute into (2.8)–(2.10); from the vanishing of the first two terms in the expansions we deduce

$$\left\{\begin{aligned} &a_1 + a_2 = 0,\\ &b_1 + b_2 + b_3 = 0,\\ &\Omega_0^{(1)} = -a_1a_2 - (b_1 + b_2),\\ &a_1a_2 = 1,\\ &a_1b_2 + a_2b_1 = 0. \end{aligned}\right. \tag{3.10}$$

Note that the equation $a_1 + a_2 = 0$ arises from the leading term of both (2.8) and (2.9). These equations imply that

$$\left\{\begin{aligned} &a_{1,2} = \pm i,\\ &b_2 = b_1,\\ &b_3 = -2b_1,\\ &\Omega_0^{(1)} = -1 - 2b_1. \end{aligned}\right. \tag{3.11}$$

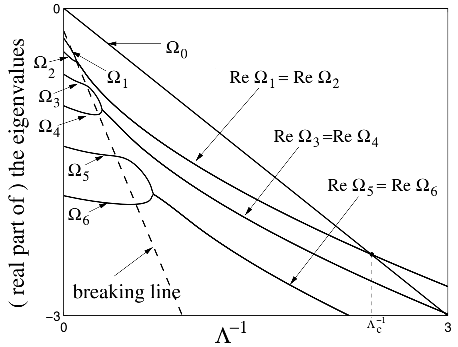


FIG. 1. *The evolution of the real parts of the first seven eigenvalues $\Omega_0, \Omega_1, \ldots, \Omega_6$ versus Λ^{-1}, assuming $L = 15$. The dashed line is the* breaking line *described in* (4.26)*, the estimate for where the eigenvalues $\Omega_1, \Omega_2, \ldots$ become complex. Λ_c^{-1} is the crossover point such that if $\Lambda^{-1} > \Lambda_c^{-1}$, the eigenvalue which has largest real part, $\Omega_{\max}$, is complex.*

Then we substitute (3.11) into the determinant (2.11). The leading term in the resulting expansion, which is $\mathcal{O}(\Lambda^{-1})$, vanishes. Requiring the $\mathcal{O}(\Lambda^{-3/2})$ term to vanish yields the claim (3.8), even though we do not yet know the b_j's.

Incidentally, if desired, the b_j's may be determined by substituting (3.4) into (3.6), yielding

$$b_1 = b_2 = -\frac{L}{2(\exp(L) - 1)} = -\frac{b_3}{2}. \tag{3.12}$$

4. A long fiber:[3] ***L* ≫ 1.** To analyze larger values of Λ^{-1}, where the expansion of section 3 loses accuracy, we need to require that $L \gg 1$. When Λ^{-1} is small, all eigenvalues are real and ordered by their index: i.e.,

$$\Omega_0 > \Omega_1 > \Omega_2 > \cdots . \tag{4.1}$$

As Λ^{-1} increases, some of the eigenvalues become complex. As this occurs we still retain the ordering

$$\text{Re } \Omega_1 \geq \text{Re } \Omega_2 > \text{Re } \Omega_3 \geq \text{Re } \Omega_4 > \text{Re } \Omega_5 > \cdots . \tag{4.2}$$

However, Ω_0 remains real,[4] and although its position in the sequence (4.2) varies with Λ^{-1}, we retain the index zero. This behavior is illustrated in Figure 1, where we have set $L = 15$.

4.1. Computation of $\mathbf{\Omega_0}$. We shall prove that, provided L is sufficiently large, the operator $\mathscr{L}$ has a real eigenvalue

$$\Omega_0 = -\Lambda^{-1} + \mathcal{O}(e^{-L}). \tag{4.3}$$

Note that this result is consistent with (3.8). Recall from section 3.2 that, in solving (2.8)–(2.10) for small nonzero Λ^{-1}, we found that $\kappa_{1,2}$ became a complex-conjugate

[3] We remind the reader that in this section, Λ and L, without bars, refer to the dimensionless parameters defined by (2.3).

[4] This statement breaks down if Λ^{-1} becomes exceedingly large, contradicting our implicit assumption that $\Lambda^{-1} \ll L$.

pair while κ_3 remained real. Thus for general Λ^{-1} we assume

$$\kappa_{1,2} = \mu \pm i\nu. \tag{4.4}$$

Then (2.8) implies

$$\kappa_3 = 1 - 2\mu. \tag{4.5}$$

We use (2.9)–(2.11) to solve for μ, ν, and Ω as follows. Dividing the middle row of (2.11) by $e^{\kappa_3 L}$, we rewrite this equation as

$$\det \begin{pmatrix} \kappa_1 & \kappa_2 & \kappa_3 \\ \kappa_1 e^{(\kappa_1 - \kappa_3)L} & \kappa_2 e^{(\kappa_2 - \kappa_3)L} & \kappa_3 \\ \Omega_0 - \kappa_1^2 & \Omega_0 - \kappa_2^2 & \Omega_0 - \kappa_3^2 \end{pmatrix} = 0. \tag{4.6}$$

Now provided

$$\mu < \frac{1}{3}, \tag{4.7}$$

we have for the 2,1- and 2,2-entries of this determinant

$$|e^{(\kappa_j - \kappa_3)L}| = e^{-(1-3\mu)L} \ll 1, \tag{4.8}$$

since $L \gg 1$. Thus neglecting these entries and recalling (4.4)–(4.5), we conclude

$$\Omega_0 = -(\mu^2 + \nu^2) + \mathcal{O}(e^{-(1-3\mu)L}). \tag{4.9}$$

On the other hand, we substitute (4.4)–(4.5) into (2.9) to find

$$\Omega_0 = -(\mu^2 + \nu^2) - 2\mu(1 - 2\mu). \tag{4.10}$$

Comparing (4.9) and (4.10), we deduce

$$\mu(1 - 2\mu) = \mathcal{O}(e^{-(1-3\mu)L}). \tag{4.11}$$

By (4.7), we have $1 - 2\mu > 1/3 > 0$, so dividing (4.11) by $1 - 2\mu$, we obtain

$$\mu = \mathcal{O}(e^{-(1-3\mu)L}). \tag{4.12}$$

Thus assumption (4.7) is consistent and, moreover,

$$\mu = \mathcal{O}(e^{-(1-3\mu)L}) = \mathcal{O}(e^{-L}). \tag{4.13}$$

Returning to (4.9), we conclude that

$$\Omega_0 = -\nu^2 + \mathcal{O}(e^{-L}). \tag{4.14}$$

But substituting (4.4)–(4.5) into (2.10) and working with (4.13), we see

$$\Lambda^{-1} = \nu^2 + \mathcal{O}(e^{-L}). \tag{4.15}$$

The claim (4.3) follows on eliminating ν from (4.14)–(4.15).

As one might expect with an exponentially small error, (4.3) agrees extremely well with the numerical results. Indeed the graph of (4.3) cannot be distinguished visually from the graph of Ω_0 in Figure 1.

Incidentally, the eigenfunction associated with Ω_0 exhibits unusual behavior for a one-dimensional eigenvalue problem—the number of its zeros or nodes changes as Λ^{-1} varies. Specifically, modulo terms that are exponentially small (outside of a boundary layer near $x = L$), the nondimensionalized eigenfunction is just $\cos(x/\sqrt{\Lambda})$.

4.2. Later eigenvalues, real case. In this subsection, we characterize a range of Λ^{-1}, as this parameter increases from 0, in which the eigenvalue Ω_n is real. Continuing the structure of the roots κ_j inherited from small Λ^{-1}, we assume

$$\kappa_{1,2} = \mu \pm i\nu \tag{4.16}$$

and invoke (2.8) to conclude

$$\kappa_3 = 1 - 2\mu. \tag{4.17}$$

We shall solve (2.12) for ν and substitute the result into (2.9)–(2.10) to obtain a parametric representation of the curve $\Omega = \Omega_n(\Lambda^{-1})$ in the Λ^{-1}, Ω-plane, with μ as the parameter. (It is not practical to solve explicitly for Ω_n as a function of Λ^{-1}.)

Let us divide the second row of the determinant (2.12) by $\exp(\mu L)$, obtaining

$$\det \begin{pmatrix} 1 & 1 & 1 \\ e^{i\nu L} & e^{-i\nu L} & e^{(1-3\mu)L} \\ \kappa_1^{-2} & \kappa_2^{-2} & \kappa_3^{-2} \end{pmatrix} = 0. \tag{4.18}$$

If

$$\mu > \frac{1}{3}, \tag{4.19}$$

then for large L we may neglect the 2, 3-entry of this determinant, so that the equation reduces to

$$e^{2i\nu L} = \frac{\kappa_2^2(\kappa_1^2 - \kappa_3^2)}{\kappa_1^2(\kappa_2^2 - \kappa_3^2)} + \mathcal{O}(e^{-(3\mu-1)L}). \tag{4.20}$$

Equation (4.20) suggests that $\nu L = \mathcal{O}(1)$ or

$$\nu = \mathcal{O}(L^{-1}). \tag{4.21}$$

Assuming this and recalling (4.19), we see that the right-hand side of (4.20) equals $1 + \mathcal{O}(L^{-1})$. Solving (4.20), we find

$$\nu = n\pi \cdot L^{-1} + \mathcal{O}(L^{-2}), \qquad n = 1, 2, \ldots, \tag{4.22}$$

confirming our assumption (4.21). On the other hand, substituting (4.16)–(4.17) into (2.8)–(2.9), we find

$$\begin{cases} \Lambda^{-1} = (1-2\mu)(\mu^2 + \nu^2), \\ \Omega_n = 3\mu^2 - 2\mu - \nu^2. \end{cases} \tag{4.23}$$

Substituting (4.22) into (4.23), we obtain the parametric representation

$$\begin{cases} \Lambda^{-1} = (1-2\mu) \cdot (\mu^2 + n^2\pi^2 L^{-2}) + \mathcal{O}(L^{-3}), \\ \Omega_n = 3\mu^2 - 2\mu - n^2\pi^2 L^{-2} + \mathcal{O}(L^{-3}). \end{cases} \tag{4.24}$$

Note from (4.24) that $\Lambda^{-1} = 0$ occurs when $\mu = 1/2$. (This may also be seen by solving (2.8)–(2.11) directly when $\Lambda^{-1} = 0$, which avoids the $\mathcal{O}(L^{-3})$ error in (4.24).) Recalling (4.19), we conclude that

$$1/3 < \mu \leq 1/2 \tag{4.25}$$

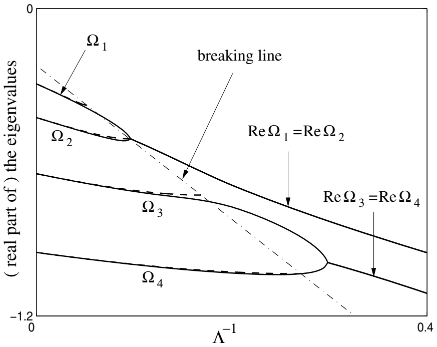


FIG. 2. *A comparison between the computations (solid lines) and the theoretical approximations given by* (4.24) *(dashed lines) of the eigenvalues* $\Omega_1, \ldots, \Omega_4$ *while they are real. The approximation terminate at the breaking line, given by* (4.26)*. The dimensionless cable length* L *is* 15.

is the relevant parameter range in (4.24). At the lower end of the range, the points (Λ^{-1}, Ω) in (4.24) all lie along the line

$$\Omega = -3\Lambda^{-1} - 2/9, \tag{4.26}$$

which we call the *breaking line*. This may be seen, avoiding the $\mathcal{O}(L^{-3})$-errors, by setting $\mu = 1/3$ in (4.23) and eliminating ν^2.

The approximations (4.24), for $n = 1, 2, 3, 4$ and for μ satisfying (4.25), are graphed in Figure 2, along with the computed eigenvalues. As the figure emphasizes, the asymptotics underlying (4.24) break down as $\mu \to 1/3$. More precisely, for μ near $1/3$, exceedingly large values of L are needed to make (4.24) accurate, and increasingly so as n becomes large.

Incidentally, note that for $\Lambda^{-1} = 0$,

$$\frac{d\,\Omega_n}{d\Lambda^{-1}} = \frac{d\,\Omega_n/d\mu}{d\Lambda^{-1}/d\mu}\Big|_{\mu=\frac{1}{2}} = -\frac{2}{1 + 4n^2\pi^2L^{-2}}, \tag{4.27}$$

which is consistent with (3.4), even without the $\mathcal{O}(L^{-3})$-errors in (4.24).

4.3. Later eigenvalues, complex case. Motivated by the numerical results, when the above asymptotics break down we look for complex eigenvalues. Let us define

$$\mu = \frac{1}{2}\ \mathrm{Re}\ (\kappa_1 + \kappa_2), \tag{4.28}$$

$$\nu = \frac{1}{2}\ \mathrm{Im}\ (\kappa_1 + \kappa_2). \tag{4.29}$$

Then

$$\kappa_{1,2} = \mu + i\nu \pm \delta, \tag{4.30}$$

where δ, possibly complex, is to be determined, and (2.8) implies that

$$\kappa_3 = 1 - 2(\mu + i\nu). \tag{4.31}$$

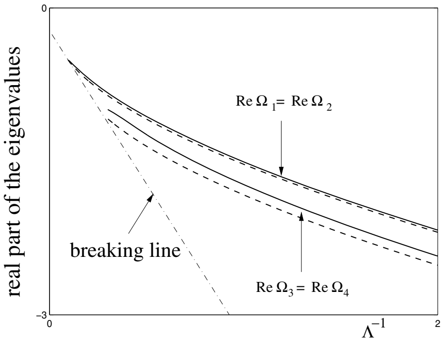


FIG. 3. *A comparison between the computations (solid lines) and the theoretical approximations given by* (4.36), (4.40) *(dashed lines) of the real parts of the eigenvalues* $\Omega_1, \ldots, \Omega_4$ *while they are complex. The dimensionless cable length* L *is* 15.

Even in the complex case, we continue to assume (4.19). Thus, asymptotically for large L, the determinant equation (2.12) may be simplified to

$$e^{2\,\delta L} = \frac{\kappa_2^2(\kappa_1^2-\kappa_3^2)}{\kappa_1^2(\kappa_2^2-\kappa_3^2)} + \mathcal{O}(e^{-(3\mu-1)L}). \tag{4.32}$$

As above, we solve (4.32) to obtain

$$\delta = in\pi\cdot L^{-1} + \mathcal{O}(L^{-2}), \quad n = 1, 2, \ldots. \tag{4.33}$$

We determine the real parameter ν from the condition that Λ^{-1} must be real: i.e., by equation (2.10)

$$\text{Im }(\kappa_1\kappa_2\kappa_3) = 0, \tag{4.34}$$

and from this we deduce that

$$\nu = \pm\Big\{\mu(3\mu-1) + n^2\pi^2L^{-2} + \mathcal{O}(L^{-3})\Big\}^{1/2}. \tag{4.35}$$

Substituting into (2.9)–(2.10), we obtain a parametric representation of Ω versus Λ^{-1}. Specifically, adjusting indices to account for the fact that (4.35) has two solutions, we find

$$\Lambda^{-1} = \mu(4\mu-1)^2 + 4\mu\cdot n^2\pi^2L^{-2} + \mathcal{O}(L^{-3}), \tag{4.36}$$

$$\Omega_{2n} = (\mu+i\nu)(3\mu+3i\nu-2) - n^2\pi^2L^{-2} + \mathcal{O}(L^{-3}), \tag{4.37}$$

$$\Omega_{2n-1} = \overline{\Omega}_{2n}, \tag{4.38}$$

where ν is given by (4.35) and

$$1/3 < \mu < \infty. \tag{4.39}$$

By substituting $\mu = 1/3$ into (4.36), (4.37) we see that these approximations terminate at the breaking line (4.26). Several of them are graphed in Figure 3, along with the computed eigenvalues. Incidentally, note that the real parts of the eigenvalues may be written more simply, without ν:

$$\text{Re }\Omega_{2n-1} = \text{Re }\Omega_{2n} = -\mu(6\mu - 1) - 4n^2\pi^2L^{-2} + \mathcal{O}(L^{-3}). \tag{4.40}$$

To explore the transition between the real and complex cases, we substitute the limiting value $\mu = 1/3$ into (4.24) and (4.36), (4.37). For the even-index eigenvalues, both the real and complex cases give same result,

$$\Lambda^{-1} = \frac{1}{27} + \frac{1}{3}\left(\frac{2n\pi}{L}\right)^2 + \mathcal{O}(L^{-3}), \tag{4.41}$$

$$\Omega_{2n} = -\frac{1}{3} - \left(\frac{2n\pi}{L}\right)^2 + \mathcal{O}(L^{-3}), \tag{4.42}$$

which of course lies on the breaking line (4.26). For odd-index eigenvalues, there is an $\mathcal{O}(L^{-2})$ jump between the results of substituting $\mu = 1/3$ into (4.24) and into (4.36), (4.37). Hints of this behavior may be seen in Figure 2—the asymptotic approximation of Ω_{2n} in the real case continues to be defined closer to the actual transition to complex eigenvalues than that of Ω_{2n-1}. Incidentally, (4.41) may be used to estimate the largest value of Λ^{-1} at which Ω_{2n-1} and Ω_{2n} are still real.

Let us relate these formulae to a result of Echebarria and Karma [4]. Invoking considerations of group velocity and bifurcation theory, those authors argue that, at the onset of instability, the complex wave number ($\kappa = \mu + i\nu$ in our notation) ought to satisfy

$$2\kappa^3 - \kappa^2 + \frac{1}{\Lambda} = 0. \tag{4.43}$$

(This relation is equivalent to their equation (56) written in our notation.) Now to leading order (4.35), (4.36) assert that

$$\nu^2 = \mu(3\mu - 1), \quad 16\mu^3 - 8\mu^2 + \mu = \Lambda^{-1}. \tag{4.44}$$

It may be checked that these two real equations are equivalent to the single complex equation (4.43). Thus we have given an independent derivation of (4.43) that does not rely on computing group velocities for waves that grow exponentially in space and that accounts explicitly for boundary conditions on a (long) finite-length cable.

5. Discussion and conclusion.

Let us summarize the asymptotic results from section 4 regarding the eigenvalues of the linear operator of (1.4), which is $\bar{w}^2\mathscr{L} = (w^2/\xi^2)\cdot\mathscr{L}$. For this task, and for the remainder of the paper, we shall undo the scaling (2.1) and (2.3) and return to the dimensional parameters. We found the following:

- Ω_0 is always real, given by

$$\Omega_0 = -\frac{\xi^2}{w\Lambda} + \mathcal{O}(e^{-wL/\xi^2}). \tag{5.1}$$

- For $n \geq 1$, Ω_n is real if Λ^{-1} is below a threshold. Equation (4.41) estimates that the threshold for Ω_{2n-1} and Ω_{2n} to become complex is

$$\frac{w^3}{27\xi^4} + \frac{w}{3}\left(\frac{2n\pi}{L}\right)^2 + \mathcal{O}(L^{-3}). \tag{5.2}$$

- In the real case, the relation between Λ^{-1} and Ω_n is given parametrically by

$$\text{(5.3)} \qquad \begin{cases} \Lambda^{-1} = \xi^{-2}(w - 2\mu\xi) \cdot (\mu^2 + n^2\pi^2\xi^2 L^{-2}) + \mathcal{O}(L^{-3}), \\ \Omega_n = 3\mu^2 - 2\mu w/\xi - n^2\pi^2\xi^2 L^{-2} + \mathcal{O}(L^{-3}), \end{cases}$$

  where $w/3\xi < \mu \leq w/2\xi$.
- In the complex case, the relation between Λ^{-1} and the real parts of Ω_{2n-1}, Ω_{2n} is given parametrically by
  (5.4)

$$\begin{cases} \Lambda^{-1} = \mu\xi^{-1}(4\mu - w/\xi)^2 + 4\mu\xi \cdot n^2\pi^2 L^{-2} + \mathcal{O}(L^{-3}), \\ \text{Re } \Omega_{2n-1} = \text{Re } \Omega_{2n} = -\mu(6\mu - w/\xi) - 4n^2\pi^2\xi^2 L^{-2} + \mathcal{O}(L^{-3}), \\ \text{Im } \Omega_{2n-1} = -\text{Im } \Omega_{2n} = 2(3\mu - w/\xi)\Big[\mu(3\mu - w/\xi) + n^2\pi^2\xi^2 L^{-2}\Big]^{1/2} + \mathcal{O}(L^{-3}), \end{cases}$$

  where $w/3\xi < \mu < \infty$.

Consider bifurcation of (1.2) from the zero solution as σ increases. An eigenvalue of the linearization crosses into the (unstable) right-half plane when $\sigma = \text{Re } \Omega_{\max}$, where $\Omega_{\max}$ is the eigenvalue of (1.4) with the (algebraically) largest real part. If Λ^{-1} is small, Ω_0 is the largest eigenvalue, its associated eigenfunction is real, and a time-independent stationary-wave solution of (1.2) appears at the bifurcation. However, Ω_0 decreases more rapidly with Λ^{-1} than later eigenvalues, which, moreover, become complex. Thus, when Λ^{-1} is sufficiently large, say $\Lambda^{-1} > \Lambda_c^{-1}$, $\Omega_{\max}$ will be complex, and time-oscillatory traveling-wave solutions will appear at the bifurcation (see [4] for more details). To estimate the crossover value Λ_c^{-1}, we consider the equation $\Omega_0 = \text{Re } \Omega_1$. Recalling (5.1), we may rewrite this equation to leading order as

$$\text{(5.5)} \qquad -\frac{\xi^2}{w\Lambda} = \text{Re } \Omega_1.$$

Substituting the first and second equations of (5.4) into the left and right sides of (5.5), respectively, we obtain the quadratic equation

$$8\mu^2 - 7(w/\xi) \cdot \mu + (w/\xi)^2 = 0$$

for the value of μ associated with the crossover. We select the root $\mu = (7+\sqrt{17})w/16\xi$ that satisfies $\mu > w/3$ and substitute into (4.36) to obtain the leading order estimate

$$\text{(5.6)} \qquad \Lambda_c^{-1} \approx \frac{71 + 17\sqrt{17}}{64} \cdot \frac{w^3}{\xi^4}.$$

By carrying $\mathcal{O}(L^{-2})$ terms in the above calculation, one may extend this estimate to next order,

$$\text{(5.7)} \qquad \Lambda_c^{-1} = \frac{w^3}{\xi^4}\left\{\frac{71 + 17\sqrt{17}}{64} + \frac{(7 + \sqrt{17})\pi^2}{2} \cdot \left(\frac{wL}{\xi^2}\right)^{-2}\right\} + \mathcal{O}(L^{-3}).$$

Figure 4 shows a comparison between the computational result for Λ_c^{-1} and theoretical approximations (5.6) and (5.7) for various (large) cable lengths L. Observe that the computational value is between the values of (5.6) and (5.7). For L smaller than shown in the figure, the graph of Λ_c^{-1} changes character. A hint of such behavior may be gleaned from the fact that the approximation (5.7) blows up as L^{-2} as L

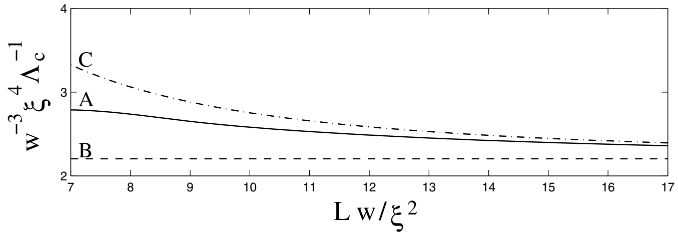


FIG. 4. *A comparison among the computational result (curve A) for Λ_c^{-1}, the theoretical leading order approximation* (5.6) *(line B), and the second order approximation* (5.7) *(curve C). The x-axis is the cable length L scaled by w/ξ^2, and y-axis is the critical value Λ_c^{-1} scaled by $w^{-3}\xi^4$. With these scalings, the curves are independent of ξ and w.*

TABLE 1

Results of two simulations: two-current model and Noble's model. w, ξ, L have units of length in cm*; Λ and Λ_c have units of inverse length in* cm^{-1}.

Name of model	w	ξ	L	Λ^{-1}	Λ_c^{-1}	Observed alternans
Two-current	0.034	0.310	25	0.206	0.011	Traveling
Noble	0.045	0.180	20	0.020	0.198	Stationary

tends to zero. We plan to investigate these phenomena more thoroughly in a future publication.

Table 1 summarizes the results of two simulations that illustrate the different behavior that occurs for $\Lambda^{-1} > \Lambda_c^{-1}$ and $\Lambda^{-1} < \Lambda_c^{-1}$. The two-current model [8], a simplified cardiac model similar in spirit to the FitzHugh–Nagumo model, is discussed in the appendix of this paper. The Noble model [9] was an early attempt to adapt a Hodgkin–Huxley-type model to cardiac cells. Because realism was attempted, this model is substantially more complicated than the two-current model. The key behavior relevant here is that the conduction-velocity curve is exceptionally flat at the critical diastolic interval, which makes $\Lambda^{-1} = 2c'/c^2$ small [3]. Simulations with the Noble model are presented in [4].

Appendix: Alternans. In this appendix we illustrate the phenomena of alternans in the context of a simple cardiac model [8], which is similar in spirit to the FitzHugh–Nagumo equation. A single heart cell in this model is described by two dimensionless functions of time, a scaled voltage v and a gate h that satisfy a set of ODEs

$$\frac{dv}{dt} = J_{\text{ion}}(v,h) + J_{\text{stim}}(t), \tag{A.1}$$

where the ionic current is

$$J_{\text{ion}}(v,h) = \frac{h}{\tau_{\text{in}}}v^2(1-v) - \frac{v}{\tau_{\text{out}}} \tag{A.2}$$

and $J_{\text{stim}}(t)$ is an external current applied repeatedly in brief pulses (cf. (A.4) below), and

$$\frac{dh}{dt} = \begin{cases} -\dfrac{h}{\tau_{\text{close}}} & \text{if } v > v_{\text{crit}}, \\[2ex] \dfrac{1-h}{\tau_{\text{open}}} & \text{if } v < v_{\text{crit}}. \end{cases} \tag{A.3}$$

Table 2
Representative values for the parameters in the two-current model.

τ_{in}	τ_{out}	τ_{open}	τ_{close}	v_{crit}	K	L
0.2ms	7ms	50ms	130ms	0.05	$0.4\ \text{cm}^2 \cdot \text{s}^{-1}$	25cm

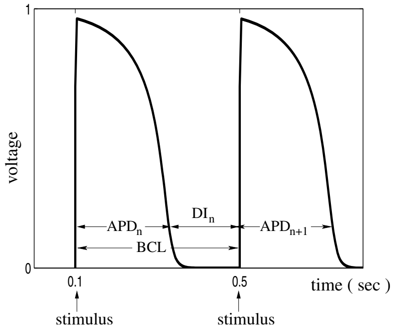


Fig. 5. *An illustration of the solution to ODE system* (A.1)–(A.3) *with parameters given in Table* 2*, assuming a periodic stimulus with period* $BCL = 400$ *ms.*

Representative values for the parameters in these equations are given in Table 2; note that K is the diffusion coefficient that appears in the PDE (A.11).

In the absence of a stimulus current, i.e., $J_{\text{stim}} = 0$, (A.1)–(A.3) have a stable equilibrium at $(v, h) = (0, 1)$. Suppose that this equilibrium is perturbed by a sequence of stimuli, applied with period B, say

$$J_{\text{stim}}(t) = \begin{cases} v_{\text{stim}}/\delta & \text{if } 0 < t < \delta \quad (\text{mod } B), \\ 0 \quad \text{otherwise}, \end{cases} \tag{A.4}$$

where $\delta \ll \tau_{\text{in}}$ and v_{stim} is not excessively small. Provided that this pacing is not too rapid, the stimuli produce action potentials as illustrated in Figure 5; i.e., each stimulus, although very brief, triggers an extended rise in the voltage, after which the voltage decays. Let the APD (A_n) and DI (D_n) be defined as in the figure; note that $A_n + D_n = B$, where B is the period or basic cycle length.

In [8] it is shown that, under the assumption

$$\tau_{\text{in}} \ll \tau_{\text{out}} \ll \tau_{\text{open}},\ \tau_{\text{close}}, \tag{A.5}$$

these variables approximately satisfy

$$A_{n+1} = F(D_n), \tag{A.6}$$

where

$$F(D_n) = \tau_{\text{close}} \ln \left\{ \frac{1 - (1 - h_{\text{min}})\, e^{-D_n/\tau_{\text{open}}}}{h_{\text{min}}} \right\}, \tag{A.7}$$

with $h_{\text{min}} = 4\tau_{\text{in}}/\tau_{\text{out}}$. Since $D_n = B - A_n$, the sequence A_n is determined by iteration of a one-dimensional map,

$$A_{n+1} = F(B - A_n). \tag{A.8}$$

Provided that B is not too small, a sequence generated by (A.8) converges to a stable fixed point $A_*(B)$. However, for B smaller than some critical value, B_{crit}, the fixed point loses its stability and we have a period-doubling bifurcation of A_n to a response called *alternans*. Let A_{crit} be the fixed point solution to (A.8) for $B = B_{\text{crit}}$, and define $D_{\text{crit}} = B_{\text{crit}} - A_{\text{crit}}$. Recognizing that $|F'(D_{\text{crit}})| = 1$ is the condition for bifurcation, we find

$$D_{\text{crit}} = \tau_{\text{open}} \ln\left\{(1 - h_{\min})\left(1 + \frac{\tau_{\text{close}}}{\tau_{\text{open}}}\right)\right\}, \tag{A.9}$$

and thus

$$A_{\text{crit}} = F(D_{\text{crit}}) = \tau_{\text{close}} \ln\left\{\frac{\tau_{\text{close}}}{(\tau_{\text{open}} + \tau_{\text{close}})h_{\min}}\right\}. \tag{A.10}$$

In a homogenized cardiac fiber, say $0 < x < L$, (A.1) is augmented by a diffusion term to obtain a PDE

$$\partial_t v = K\partial_{xx} v + J_{\text{ion}}(v, h) + J_{\text{stim}}(x, t); \tag{A.11}$$

equation (A.3) does not acquire any additional terms, but the t-derivative must be reinterpreted as a partial derivative. The stimulus current is applied locally near one end of the fiber and vanishes elsewhere. No-flux boundary conditions are imposed at both ends of the fiber:

$$\partial_x v(0, t) = \partial_x v(L, t) = 0. \tag{A.12}$$

The action potentials stimulated near $x = 0$ propagate along the fiber. The traveling speed c, or the conduction velocity (CV) of a periodic wave train, depends on the DI D,

$$c = c(D) \approx \sqrt{\frac{K \cdot h_{\text{init}}}{2\tau_{\text{in}}}}\left(1 - \frac{3h_{\min}}{4h_{\text{init}}}\right), \tag{A.13}$$

where

$$h_{\text{init}}(D) = 1 - (1 - h_{\min})e^{-D/\tau_{\text{open}}}. \tag{A.14}$$

If the BCL of the stimuli is sufficiently small (i.e., if the pacing frequency is high enough), alternans will appear along the cable.

Let $A_k(x)$ denote the duration of the kth action potential at position x. Assume that $A_k(x)$ has the form (1.1) in section 1. Echebarria and Karma [4] derived the approximate equation (1.2) to describe the evolution of the amplitude of alternans. The parameters Λ, w, ξ that appear in (1.2) are estimated by

$$\begin{cases} \Lambda^{-1} = c'/c^2, \\ w = 2K/c, \\ \xi = \sqrt{K \cdot A_{\text{crit}}}, \end{cases} \tag{A.15}$$

where c and $c' = \frac{dc}{dD}$ in (A.15) are both evaluated at the critical value $D = D_{\text{crit}}$. Note that D_{crit} and A_{crit} are given by (A.9), (A.10). Evaluating $c(D_{\text{crit}})$ by substituting

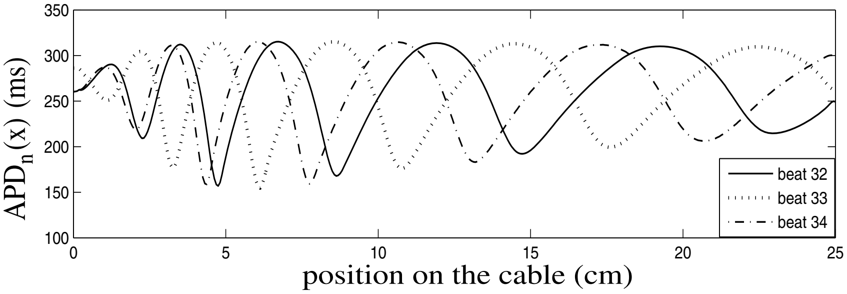


FIG. 6. *Simulation of the PDE* (A.11) *with parameters given in the Table* 2*, assuming a periodic stimulus with period $BCL = 343$ ms. The x-axis is the position on the cable, the y-axis is $A_n(x)$ for the values of n listed above. As predicted by using* (5.7)*, the pattern is propagating.*

the parameters of Table 2 into (A.13), we compute w and ξ as given in Table 1. On the other hand, by differentiating (A.13), we obtain

$$c' = \frac{dc}{dD} = \frac{1 - h_{\min}}{\tau_{\text{open}}}\sqrt{\frac{K}{2\tau_{\text{in}}}} \cdot \left(\frac{1}{2}h_{\text{init}}^{-1/2} + \frac{3}{8}h_{\min}h_{\text{init}}^{-3/2}\right)\ e^{-D/\tau_{\text{open}}}, \tag{A.16}$$

and after substituting we find Λ^{-1} as given in Table 1. The simulation shown in Figure 6 were performed on a cable of length 25 cm, which corresponds to a scaled dimensionless length of

$$\bar{\bar{L}} = wL/\xi^2 = 8.84. \tag{A.17}$$

For this length, the (dimensionless) critical value Λ_c^{-1}(computed numerically—curve A in Figure 4) equals 2.67, which in dimensional units gives the value listed in Table 1. Since $\Lambda^{-1} > \Lambda_c^{-1}$, the modulation equation predicts that alternans will appear in traveling patterns. The behavior is observed in the simulation of Figure 6, which shows $A_n(x)$ for several beats about halfway through the transient to steady state.

REFERENCES

[1] American Heart Association website: http://www.americanheart.org/.

[2] D. R. Chialvo, D. C. Michaels, and J. Jalife, *Supernormal excitability as a mechanism of chaotic dynamics of activation in cardiac Purkinje fibers*, Circ. Res., 66 (1990), pp. 525–545.

[3] B. Echebarria and A. Karma, *Instability and spatiotemporal dynamics of alternans in paced cardiac tissue*, Phys. Rev. Lett., 88 (2002), paper 208101.

[4] B. Echebarria and A. Karma, *Amplitude-equation approach to spatiotemporal dynamics of cardiac alternans*, Phys. Rev. E, 76 (2007), paper 051911.

[5] A. Garfinkel, Y.-H. Kim, O. Voroshilovsky, Z. Qu, J. R. Kil, M. Hyoung, H. S. Karagueuzian, J. N. Weiss, and P.-S. Chen, *Preventing ventricular fibrillation by flattening cardiac restitution*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 6061–6066.

[6] R. F. Gilmour, Jr. and D. R. Chialvo, *Electrical restitution, critical mass, and the riddle of fibrillation*, J. Cardiovasc. Electrophysiol., 10 (1999), pp. 1087–1089.

[7] Heart Rhythm Society, *Sudden cardiac death*, online article at http://www.hrspatients.org/patients/heart_disorders/cardiac_arrest/.

[8] C. C. Mitchell and D. G. Schaeffer, *A two-current model for the dynamics of the cardiac membrane*, Bull. Math. Biol., 65 (2003), pp. 767–793.

[9] D. Noble, *A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pace-maker potentials*, J. Physiol., 160 (1962), pp. 317–352.

[10] A. V. Panfilov, *Spiral breakup as a model of ventricular fibrillation*, Chaos, 8 (1998), pp. 57–64.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 720–742

TRANSONIC SHOCK FORMATION IN A RAREFACTION RIEMANN PROBLEM FOR THE 2D COMPRESSIBLE EULER EQUATIONS[*]

JAMES GLIMM[†], XIAOMEI JI[‡], JIEQUAN LI[§], XIAOLIN LI[¶], PENG ZHANG[‖], TONG ZHANG[**], AND YUXI ZHENG[††]

Abstract. It is perhaps surprising for a shock wave to exist in the solution of a rarefaction Riemann problem for the compressible Euler equations in two space dimensions. We present numerical evidence and generalized characteristic analysis to establish the existence of a shock wave in such a 2D Riemann problem, defined by the interaction of four rarefaction waves. We consider both the customary configuration of waves at the right angle and also an oblique configuration for the rarefaction waves. Two distinct mechanisms for the formation of a shock wave are discovered as the angle between the waves is varied.

Key words. 2D Riemann problem, gas dynamics, shock waves, generalized characteristic analysis, front tracking method

AMS subject classifications. Primary, 35L65, 35J70, 35R35; Secondary, 35J65

[*] Received by the editors October 23, 2007; accepted for publication (in revised form) August 18, 2008; published electronically December 17, 2008. Computations were performed on the Galaxy computer and seawulf computer at Stony Brook and the Stony Brook-Brookhaven New York Blue computer. This research utilized resources from the Galaxy computer at Stony Brook and the New York Center for Computational Sciences at Stony Brook University/Brookhaven National Laboratory, which is supported by the U.S. Department of Energy under contract DE-AC02-98CH10886 and by the State of New York. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siap/69-3/70632.html

[†] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, and Computational Science Center, Brookhaven National Laboratory, Upton, NY 11973-5000 (glimm@ams.sunysb.edu). The research of this author was partially supported by the DOE and ARO with ID DEFG5206NA26208 and ID W911NF0510413.

[‡] Corresponding author. Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600 (xji@ams.sunysb.edu). The research of this author was partially supported by The Project sponsored by SRF for ROCS, SEM grant 2004176 when she worked in Beijing University of Technology, China before 8/2005, and was supported by the AMS Department, Stony Brook University thereafter.

[§] Department of Mathematics, Capital Normal University, Beijing 100037, People's Republic of China (jiequan@mail.cnu.edu.cn). The research of this author was partially supported by the Key Program from Beijing Educational Commission grant KZ200510028018, the Program for New Century Excellent Talents in University (NCET), and Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR-IHLB).

[¶] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600 (linli@ams.sunysb.edu). The research of this author was partially supported by the DOE with ID DEFC0206ER25770.

[‖] Department of Mathematics, Capital Normal University, Beijing 100037, People's Republic of China (pzhang@mail.cnu.edu.cn). The research of this author was supported by the Key Program of Natural Science Foundation from Beijing Municipality grant 4051002.

[**] Institute of Mathematics, Chinese Academy of Mathematics and System Sciences, Beijing 100080, People's Republic of China (tzhang@math.ac.cn). The research of this author was partially supported by The Project sponsored by SRF for ROCS, SEM grant 2004176, the Key Program from Beijing Educational Committee grant KZ200510028018, and NSFC 10671120.

[††] Department of Mathematics, Penn State University, State College, PA 16802 (yzheng@math.psu.edu). The research of this author was partially supported by NSF-DMS-0305497, 0305114, and 0603859.

DOI. 10.1137/07070632X

1. Introduction.

Shock reflection in gas dynamics has long been an open problem. Two-dimensional (2D) Riemann problems have been proposed for the compressible Euler equations as a general approach to the shock reflection problem [20]. Numerical simulations for this type of data have been performed by Chang, Chen, and Yang [1, 2], Schulz-Rinne, Collins, and Glaz [17], Lax and Liu [11, 16], Kurganov and Tadmor [10], and Li, Zhang, and Yang [13], among others. General patterns of shock reflections have been revealed, some cases of which are accessible to analytical treatment.

The initial data of a general Riemann problem is constant along radial directions from an origin and is piecewise constant as a function of angle. We consider the special case with initial data of piecewise constant solutions joined by four forward rarefaction waves; see section 2. For this case, the solution was conjectured to be continuous; see [1, 2, 17, 11, 13]. From the point of view of physically motivated wave interactions, arbitrary angles between waves may be considered, and special solutions (stationary wave interactions) in general will occur at angles other than 90°; see [8]. From the point of view of defining a Riemann solution for a finite difference mesh, we might consider a variety of meshes with different angles between the cell edges. In accordance with both points of view, we consider the oblique four-wave Riemann problem. We perform refined numerical experiments, using the FronTier code developed at the AMS department of SUNY Stony Brook, and obtain resolved numerical solutions. This code uses a five point vectorized split MUSCL scheme [5] as a shock capturing algorithm. It is second order accurate for smooth solutions and first order accurate near shock waves. We solve the full compressible Euler equations in the original x, y, t coordinates, not in self-similar coordinates, so the numerics are actually very well documented in previous literature [5] and [6]. Our main result is the existence of a shock wave, established numerically by several different criteria, for a 2D Riemann problem with four rarefaction waves in both the 90° case and the oblique case. The possibility of shock formation indicates the deep sophistication of this seemingly easy problem. We formulate plausible structures for the solution via the method of generalized characteristic analysis (i.e., the analysis of characteristics, shocks, and sonic curves or the law of causality). In section 2, we formulate the problem under study and discuss an algorithm for the construction of characteristics in the numerical solutions. The existence of shock waves is established by multiple criteria used in our numerical studies to indicate the presence of shock waves. Specifically, we consider

1. plots of density and pressure on a curve through the shocks;
2. nontangential termination of characteristics at the shock front;
3. convergence of characteristics of the same family at the local shock front;
4. pattern recognition software for automated shock wave detection;
5. stability of the above criteria under mesh refinement.

In section 3, we summarize numerical results for several cases. In section 4, we study the 90° case, and in section 5, we consider the oblique case and numerically prove the convergence of very weak shock. In section 6, we present related evidence for shock formation in the case of two backward and two forward rarefaction waves. In section 7, we discuss the physical mechanism that leads to the shock formation in the present problem and summarize results testing the stability of the numerical solutions.

2. The problem formulation and its characteristic curves. We consider the Euler equations

$$\begin{aligned} &\rho_t + \nabla \cdot (\rho U) = 0, \\ &(\rho U)_t + \nabla \cdot (\rho U \otimes U) + \nabla p = 0, \\ &(\rho E)_t + \nabla \cdot ((\rho E + p)U) = 0 \end{aligned} \tag{2.1}$$

for the variables (ρ, U, E), where ρ is the density, $U = (u, v)$ is the velocity, p is the pressure, $E = \frac{1}{2}|U|^2 + e$ is the specific total energy, and e is the specific internal energy. We consider a polytropic gas with pressure p defined by the equation

$$e = \frac{p}{(\gamma - 1)\rho} \ .$$

For more details, see the books by Li, Zhang, and Yang [13] or Zheng [21].

We solve the full compressible flow equations (2.1) in the original x, y, t coordinates. Our numerical studies are based on (2.1), using the MUSCL algorithm [5] as implemented in the FronTier code. This code uses a five point vectorized split MUSCL scheme [5] as a shock capturing algorithm. It is the second order accurate for smooth solutions and the first order accurate near shock waves. Both the MUSCL algorithm and FronTier code have been extensively verified for shock capturing simulations, for example in [5] and [6]. Shock jump conditions for (2.1) can be found in standard textbooks, for example in [3]. The numerical verification of these jump conditions is addressed in section 5.

Because (2.1) and the initial data are both self-similar, the solution is also, and we introduce the self-similar coordinate system $(\xi, \eta) = (\frac{x-x_0}{t}, \frac{y-y_0}{t})$ centered at the point (x_0, y_0). In these coordinates, the system (2.1) takes the form

$$\begin{aligned} &-\xi\rho_\xi - \eta\rho_\eta + (\rho u)_\xi + (\rho v)_\eta = 0, \\ &-\xi(\rho u)_\xi - \eta(\rho u)_\eta + (\rho u^2 + p)_\xi + (\rho uv)_\eta = 0, \\ &-\xi(\rho v)_\xi - \eta(\rho v)_\eta + (\rho v^2 + p)_\eta + (\rho uv)_\xi = 0, \\ &-\xi(\rho E)_\xi - \eta(\rho E)_\eta + (\rho u(E + \tfrac{p}{\rho}))_\xi + (\rho v(E + \tfrac{p}{\rho}))_\eta = 0. \end{aligned} \tag{2.2}$$

Let $\eta = \eta(\xi)$ be a smooth discontinuity with limit states (ρ_1, u_1, v_1, p_1) and (ρ_0, u_0, v_0, p_0) on both sides. The Rankine–Hugoniot relation for (2.2) is derived in [13, pp. 218–219]. By definition, Riemann initial data is constant along radial directions from an origin (x_0, y_0) and piecewise constant as a function of angle. The initial data for (2.1) become boundary data at infinity for (2.2). We use the self-similar formulation (2.2) for the analysis of numerical solutions of (2.1). We specialize to a four-rarefaction wave Riemann problem.

As a special case we consider first the case of four rectangularly oriented waves, representing boundary conditions at infinity for the self-similar Euler equations (2.2) satisfying conditions of four forward rarefaction waves, denoted configuration A in [13, p. 237]. We next consider the case of four constant states joined by forward rarefaction waves that form angles different from 90°, as in Figure 1. Such a problem is called *an oblique four-wave Riemann problem*, in contrast to the rectangular four-wave Riemann problem discussed in [20]. Our initial data is located as indicated in Figure 1 in the initial plane:

$$(\rho, u, v, p) = (\rho_i, u_i, v_i, p_i), \quad i = 1, 2, 3, 4. \tag{2.3}$$

Let R_{ij} denote the forward rarefaction wave, which is a 1D rarefaction wave, connecting contiguously constant states (ρ_i, u_i, v_i, p_i) and (ρ_j, u_j, v_j, p_j). R_{12} is parallel to the positive y-axis, and R_{41} is parallel to the positive x-axis as before, but the angle between R_{23} and the negative x-axis is allowed to be a variable θ in $(0, \pi/4)$. To

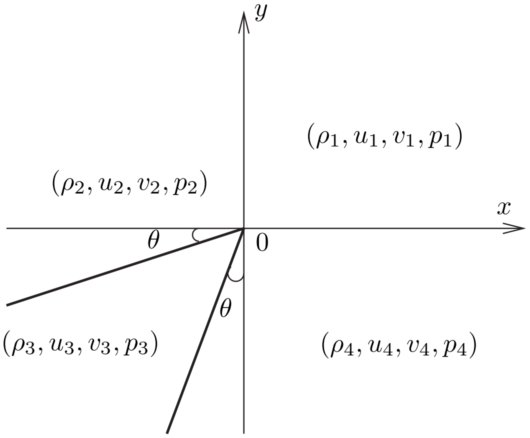

y
(ρ_1, u_1, v_1, p_1)
(ρ_2, u_2, v_2, p_2)
x
θ
0
θ
(ρ_3, u_3, v_3, p_3)
(ρ_4, u_4, v_4, p_4)

FIG. 1. *The initial data for an oblique four-wave Riemann problem.*

simplify the analysis, we impose symmetry about the line $x = y$. We choose the angle between R_{34} and the negative y-axis to be the same θ, so that the angle between R_{23} and R_{34} is equal to $\frac{\pi}{2} - 2\theta$. Let w represent the velocity component that is perpendicular to the line of discontinuity, and w' represent the velocity component parallel to it. At an interface $(i, j) \in \{(1, 2), (2, 3), (3, 4), (4, 1)\}$, a forward planar rarefaction wave R_{ij} is described by the formula in [11],

$$w_i - w_j = \frac{2\gamma^{\frac{1}{2}}}{\gamma - 1}\left(\left(\frac{p_i}{\rho_i}\right)^{\frac{1}{2}} - \left(\frac{p_j}{\rho_j}\right)^{\frac{1}{2}}\right) , \qquad w'_i = w'_j \ , \frac{p_i}{p_j} = \left(\frac{\rho_i}{\rho_j}\right)^{\gamma} . \tag{2.4}$$

For each R_{ij}, the compatibility conditions derived from (2.4), using the normal and tangential components of $u_i, v_j, i, j = 1, 2, 3, 4$, along R_{ij}, are

$$(\rho_3^{(\gamma-1)/2} - \rho_4^{(\gamma-1)/2}) \cos\theta - (\rho_2^{(\gamma-1)/2} - \rho_3^{(\gamma-1)/2}) \sin\theta + (\rho_1^{(\gamma-1)/2} - \rho_2^{(\gamma-1)/2}) = 0; \tag{2.5}$$

$$(\rho_2^{(\gamma-1)/2} - \rho_3^{(\gamma-1)/2}) \cos\theta - (\rho_3^{(\gamma-1)/2} - \rho_4^{(\gamma-1)/2}) \sin\theta - (\rho_1^{(\gamma-1)/2} - \rho_4^{(\gamma-1)/2}) = 0. \tag{2.6}$$

We limit ourselves to the initially symmetric case $\rho_2 = \rho_4$ and $u_1 = v_1$. Then the two compatibility conditions merge to yield

$$\rho_2^{(\gamma-1)/2}(\cos\theta + \sin\theta + 1) = \rho_1^{(\gamma-1)/2} + \rho_3^{(\gamma-1)/2}(\sin\theta + \cos\theta). \tag{2.7}$$

For any fixed $\rho_1, p_1, u_1, v_1, \rho_3$, and θ, we find ρ_2 from the compatibility condition (2.7) and other initial values from (2.4) and symmetry. We consider a fixed polytropic index $\gamma = 1.4$. The computational domain is a square $[0, 1] \times [0, 1]$. We perform numerical experiments with varying Riemann initial data. We draw both families of (pseudo) characteristic curves corresponding to $\lambda_\pm$ in [13],

$$\frac{d\eta}{d\xi} = \lambda_\pm(\xi, \eta) \equiv \frac{(u - \xi)(v - \eta) \pm c[(u - \xi)^2 + (v - \eta)^2 - c^2]^{1/2}}{(u - \xi)^2 - c^2} , \tag{2.8}$$

where c is the sonic speed, $\xi = \frac{x - x_0}{T_0}$, $\eta = \frac{y - y_0}{T_0}$, T_0 is fixed, and $x_0 = y_0 = 0.5$ is the center of the computational domain. By the definition in [13], the pseudo-Mach number is

$$M = \frac{[(u - \xi)^2 + (v - \eta)^2]^{1/2}}{c} . \tag{2.9}$$

The $M = 1$ contour, as understood here, indicates both sonic points and shock points, where M jumps from a value less than 1 to a value greater than 1. The sonic curve is thus a subset of the $M = 1$ contour line. We notice $\lambda_+ = \lambda_-$ on the sonic curve.

We discuss the algorithm for characteristics. The characteristic curves starting at the top boundary of the rectangular domain belong to the family λ_+, while the characteristic curves from the right boundary of the rectangular domain belong to λ_-. To draw the $\lambda_\pm$ characteristics, we assume that the numerical solution of the Euler equations is defined on a rectangular grid. We extend this solution to the entire computational domain for a discrete time $t = T_0$, using linear interpolation. Thus $\lambda_\pm$ become globally defined functions. Starting at the the right boundary, we solve for λ_- to obtain the pseudocharacteristic curves, using the Runge–Kutta scheme. The solution for λ_- is continued up to the sonic curve. For the reflected characteristics $\lambda_\pm(\xi, \eta)$ at the sonic curve, we repeat the above processes. Since these characteristics are reflections of the previously constructed family, we use bilinear interpolation to obtain initial states at the point on the $M = 1$ contour where an incoming characteristic has terminated. We solved all singularities in the characteristic equations (2.8) numerically. Since the main point of this paper is to establish the existence of a shock wave, we list here criteria that we use for this purpose.

The most sensitive of our measures for existence of a shock wave is the fact that a shock will appear when the two families of $\lambda_\pm$ characteristics are not parallel at the $M = 1$ contour line. The existence of a point on the $M = 1$ contour line with λ_+ not parallel to λ_- contradicts (2.8) if the point is a sonic point, i.e., a point at which the solution is continuous. We thus plot $\lambda_- - \lambda_+$ versus the angle around the $M = 1$ contour, where the shock is identified as the locus of points on the contour with $\lambda_- - \lambda_+ > 0$. The end points of the shock are identified relative to figures showing characteristics.

A second test for existence of a shock is to show convergence of nearby characteristics of a common family, so that they meet on the $M = 1$ contour.

As a third test, we plot ρ versus distance along a streamline. By definition in [13], pseudostream curves satisfy $\frac{d\eta}{d\xi} = \lambda_0 = \frac{v-\eta}{u-\xi}$. The solution to (2.1) is called a compression wave if $(1, u, v) \cdot (\rho_t, \rho_x, \rho_y) > 0$; otherwise it is called an expansion wave. We notice the fact

$$(2.10) \qquad (u - \xi, v - \eta) \cdot (\partial_\xi, \partial_\eta) = t(1, u, v) \cdot (\partial_t, \partial_x, \partial_y) = t\frac{d}{dt},$$

where $\frac{d}{dt}$ is evaluated along the trajectories of gas particles in (t, x, y)-space. All pseudostream curves point to the center of the subsonic domain. Moreover, we have

$$\begin{aligned}
\frac{d\rho(\xi,\eta)}{dt} &= \frac{\partial\rho}{\partial\xi} \cdot \frac{\partial\xi}{\partial t} + \frac{\partial\rho}{\partial\eta} \cdot \frac{\partial\eta}{\partial t} \\
&= \frac{\partial\rho}{\partial\xi} \cdot \left(\frac{\partial\rho}{\partial x}\frac{dx}{dt} - \frac{x}{t^2}\right) + \frac{\partial\rho}{\partial\eta} \cdot \left(\frac{\partial\rho}{\partial y}\frac{dy}{dt} - \frac{y}{t^2}\right) \\
&= \frac{\partial\rho}{\partial\xi} \cdot \left(\frac{u}{t} - \frac{\xi}{t}\right) + \frac{\partial\rho}{\partial\eta} \cdot \left(\frac{v}{t} - \frac{\eta}{t}\right) \\
&= \frac{1}{t}\left(u - \xi, v - \eta\right) \cdot (\rho_\xi, \rho_\eta) \\
&= \frac{1}{t}\frac{d\rho(\xi,\eta)}{ds}\ ,
\end{aligned}$$

Table 1
Table of simulation cases studied in this paper.

Three cases	90°	Weakly oblique rarefaction	Oblique rarefaction
Initial conditions	$\rho_1 = 1.0, p_1 = 0.444, u_1 = v_1 = 0.0, \rho_3 = 0.15$		
θ values	$\theta = 0°$	$6° \leq \theta \leq 8°$	$\theta > 8°$
Shock case	Weak shock	Weak shock	Shock
Mesh size $\Delta x = \Delta y$	$\frac{1}{3200}, \frac{1}{1600}$	$\frac{1}{5600}, \frac{1}{3200}, \frac{1}{1600}, \frac{1}{800}$	$\frac{1}{800}$

where in the last term, $\frac{d\rho(\xi,\eta)}{ds}$ is the directional derivative of the density ρ along the pseudostream curve. According to (2.10), a positive value for this derivative indicates a compression or a shock, and a jump indicates a shock.

We also plot pressure along a ray passing through the $M = 1$ contour. A sharp jump in pressure is a sign of a possible shock.

Finally, we use the wave detection filter, or automated shock detection capability in the FronTier code, to locate shock waves in [19]. This software examines numerical discontinuities in the solutions and determines whether they correspond to a particular type of traveling wave. Such discontinuities are organized into curves, which then correspond to the location of a shock wave. The shock jump conditions are included in the Rankine–Hugoniot relation. These jump conditions are verified to confirm that the shock waves detected by the wave filter indeed satisfy the required conditions. Numerical solutions show stability of the above shock criteria under mesh refinement.

Our major point is the consistency of these tests, each reaching the same conclusion: shocks exist in the interior nonconstant domain in the 2D rarefaction Riemann problems studied here. We find two distinct mechanisms for shock formation. For the 90° case, the interaction of two rarefaction waves of the same family and parallel at infinity leads to a pressure drop larger than that due to either taken singly. Thus the interaction seems to "overrarefy," leading to low pressure states incompatible with pressures given at infinity due to the same rarefactions considered individually. A shock wave results from the joining of these high and lower pressure regions. It is the interaction of rarefaction R_{41}, R_{23}, R_{34}, and R_{12}, including the interaction of characteristics from constant states adjacent to them, which produce this result. A second mechanism arises in the case of a (sufficiently) large oblique angle between the rarefaction waves. In this case, the rarefactions R_{41} and R_{12} are reflected from the sonic curve. The sonic curve has extended ears to facilitate this reflection. The reflected wave becomes a compression and breaks into a shock at the $M = 1$ contour, at points that would otherwise be sonic but actually lie on a shock front.

3. Numerical results. We summarize our further refined numerical results in this section. We fix the computational time $T_0 = 0.375$ and let the initial values be p_1, ρ_1, $u_1 = v_1 = 0$, ρ_3 and determine all initial values with different θ. We list the cases to be considered in Table 1. We denote by $C_i(u_i, v_i, c_i)$, $i = 1, 2, 3, 4$, the sonic circles of constant states i, $i = 1, 2, 3, 4$, where u_i, v_i, $i = 1, 2, 3, 4$, are initial velocities and $c_i = \left(\frac{\gamma p_i}{\rho_i}\right)^{\frac{1}{2}}$, $i = 1, 2, 3, 4$, are initial sound speeds.

A. Weak shock case for $\theta = 0$ (90° case). The 90° case is shown in Figure 2 with results consistent with on both coarse and refined grids. The λ_--characteristic lying at the upper boundary of R_{41} coming from infinity penetrates the rarefaction wave $R_{12}(CQ)$ and continues through the constant state 2 (QR) and the rarefaction wave $R_{23}(RF)$, reaching the constant state 3, and meets the sonic circle $C_3(u_3, v_3, c_3)$ tangentially at A. See the strict proof in [13, p. 238]. The bottom boundary of R_{23} (FT) hits the $M = 1$ contour at T. The top boundary of $R_{23}(RV)$ hits the

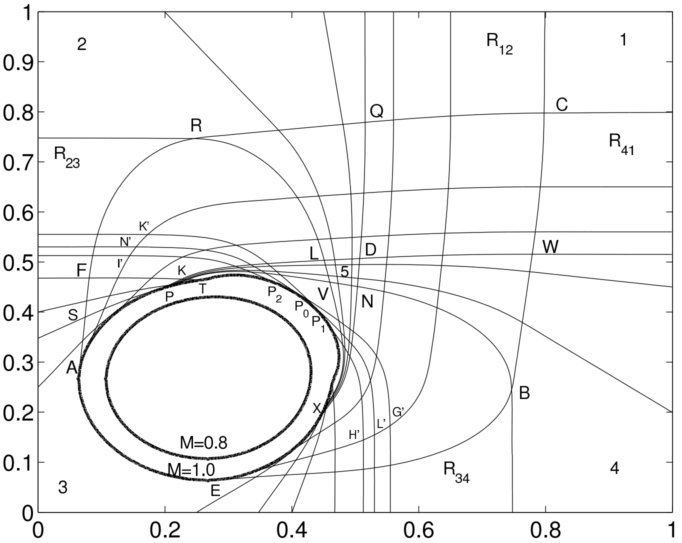


Fig. 2. *Case A: Some pseudocharacteristic curves (light) and Mach number contours (bold) marked with $M = 1.0$ and $M = 0.8$ at $\theta = 0$.*

$M = 1$ contour at X, and a weak shock appears on the larger arc AE (toward the first quadrant), where A, E are symmetric points relative to reflection about the axis $\xi = \eta$. The smaller arc AE (toward the third quadrant) is an arc of the sonic circle C_3. Numerical evidence supports weak shocks on both sides of the sonic circle. The shocks and the numerical evidence for them are stronger on the side nearer the origin. We will discuss these points in detail later.

B. Weak shock case: numerical solutions with $\theta \in [6°, 8°]$. We obtain weak shock cases for $\theta \in [6°, 8°]$. See Figures 3 and 17 for the cases $\theta = 6.5°$ and $\theta = 8°$. In Figure 3, both the upper boundaries FS of R_{41} and GS of R_{23} are parallel, tangential to the sonic curve at S. SA is tangential to the sonic circle C_3 at A, which is a reflection of the λ_+-characteristic curve GS. SK, which is the reflection of the λ_--characteristic curve FS, terminates on a shock. The weak shock appears on the arc $AT \bigcup BR \bigcup UE$, where A, E are symmetric points regarding the axis $\xi = \eta$, and the smaller arc AE is the arc of sonic circle C_3.

C. Strong shock case: numerical solutions with $\theta > 8°$. We increase the value of θ, and we observe a shock wave which is sharply defined and with little numerical oscillation. The shock wave lies in the interior, nonconstant domain between R_{23} and R_{34} and constant states adjacent to them whose structure is similar to case B. See Figure 4. Furthermore, we find that the strength of the shock wave becomes stronger as we increase θ or decrease ρ_3 while keeping the other parameters constant. The numerical oscillations between R_{23} and R_{34} attenuate or even vanish as the strength of the shock wave intensifies. In summary, numerical solutions show the existence of shock waves in the interaction of four rarefaction waves and constant states.

4. Shock formation in the 90° case. A constructive analysis of some of the major ideas of this paper is found in [7], where we study the Riemann problem for the Hamilton–Jacobi equations as a simpler problem, developing a number of ideas needed here. These equations can be regarded as a generalization of Burgers' equation to high dimensions. The analysis there follows a constructive point of view and thus

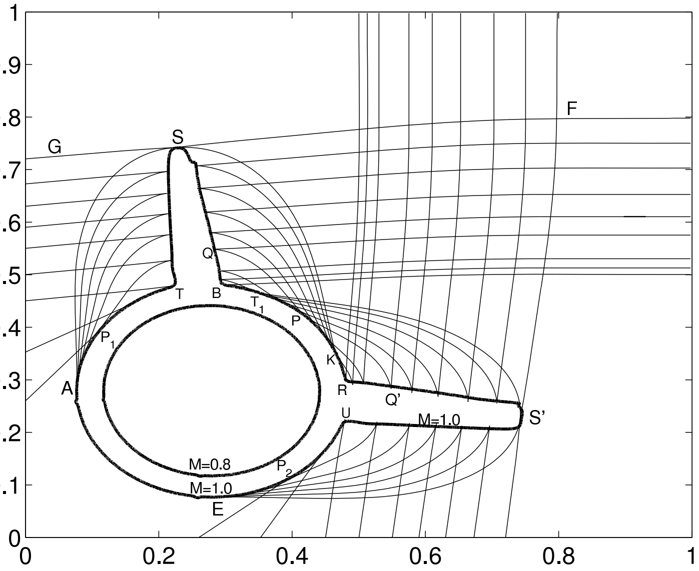


FIG. 3. *Case B: Some pseudocharacteristic curves (light) with $\theta = 6.5°$ and Mach number contours (bold) marked with $M = 1.0$ and $M = 0.8$.*

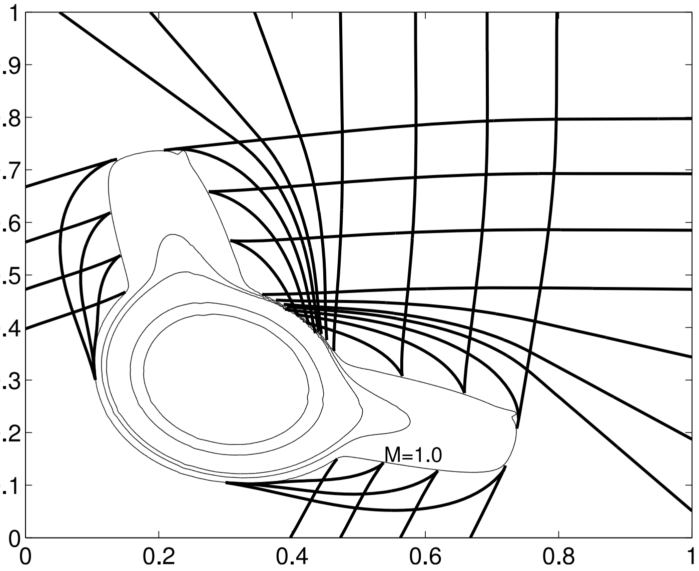


FIG. 4. *Case C: Some pseudocharacteristic curves (bold) with $\theta = 22.5°$ and Mach number contours (light) with the sonic curve labeled $M = 1.0$.*

emphasizes ideas such as generalized characteristics, the propagation of the Riemann solution inward from data located at infinity, and a sonic curve as discussed in the present paper. We discuss in this section the shock formation in the 90° case. We analyze the mechanism for shock formation caused by the interaction of rarefaction waves R_{12}, R_{23}, R_{34}, and R_{41}, including the interaction of R_{41} and characteristics from constant state $3, 4, 5$ adjacent to them. We use numerical methods and generalized characteristic analysis.

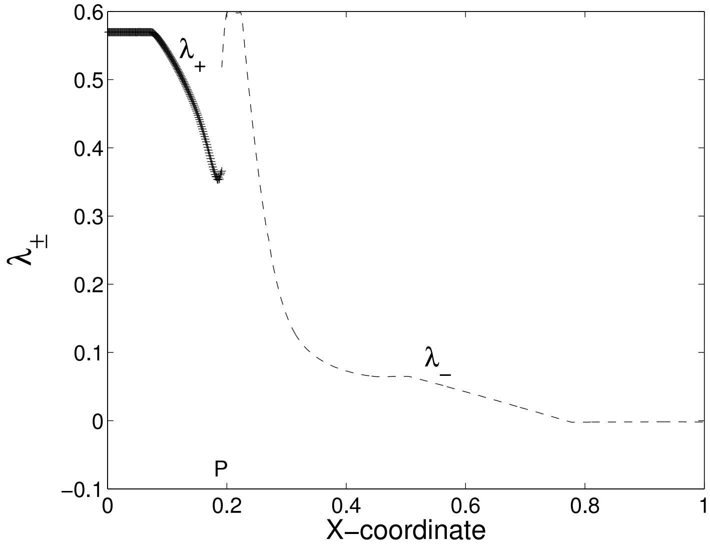


Fig. 5. *The characteristics* $\lambda_+ = \frac{d\eta}{d\xi}$ *along* SP *and* $\lambda_- = \frac{d\eta}{d\xi}$ *along* WP *meet at* P*. Note that* $\lambda_+ \neq \lambda_-$ *at the common point* P*.*

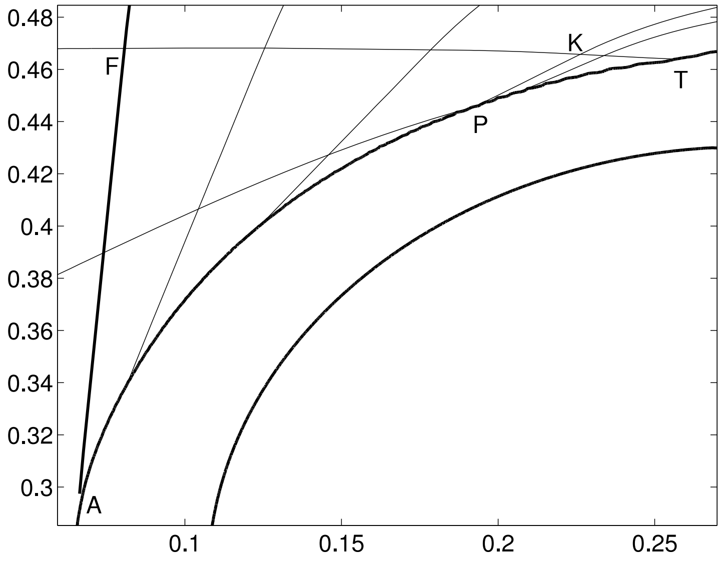


Fig. 6. *Enlarged view from Figure* 2 *in the* 90° *case shows the nonparallel termination of characteristics on the* $M = 1.0$ *contour (outer curved arc) and shock existence at the point* P*. The lower curved arc is the* $M = 0.8$ *contour.*

4.1. A numerical study of the 90° case. In Figure 2, the characteristic WP along the bottom boundary of R_{41} meets the characteristics SP from constant state 3 at P. The intersection point P is located on the $M = 1$ contour. If P were a sonic point, then we would have $\lambda_+(P) = \lambda_-(P)$. However, we show numerically in Figures 5 and 6 that P cannot be a sonic point because the characteristics are not parallel at P. The numerical results in Figure 5 show that at the common point P of the

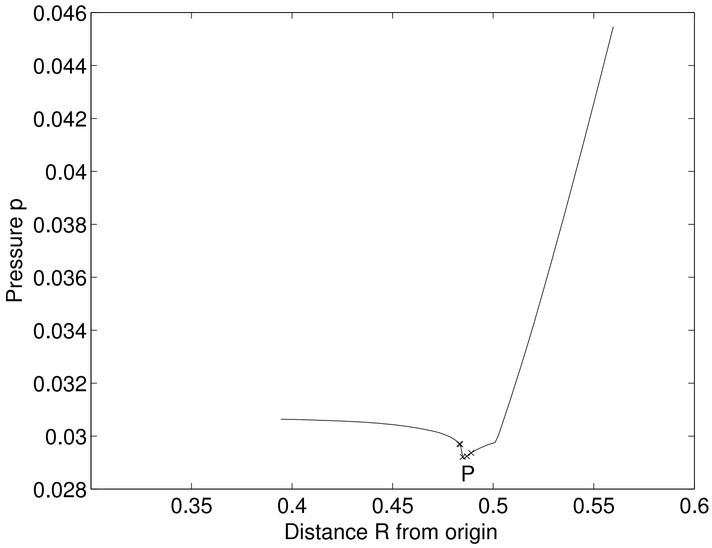


FIG. 7. *Pressure vs. distance along OP for the $90°$ case. The downward jump at the point P is a shock front. The crosses mark cell center locations near the shock front.*

two characteristic curves, $\lambda_+(P) < 0.4$, $\lambda_-(P) > 0.5$, and $\lambda_+(P) \neq \lambda_-(P)$, indicating that P is not sonic. Thus the termination of the λ_+ characteristic and the beginning of the λ_- characteristic must lie on a shock curve. We have two characteristics $N'P_0$ and $L'P_0$ meet at P_0 on $\xi = \eta$, but they are not parallel. Similarly, $G'P_2$ and $I'P_2$ meet at P_2 nontangentially, $H'P_1$ and $K'P_1$ meet at P_1 nontangentially. P_0, P_1, P_2 are located on a shock front. An enlarged view of the nontangential and nonparallel termination of the characteristic curves at the shock front is shown in Figure 6. We plot pressure versus the distance $R = (x^2 + y^2)^{\frac{1}{2}}$ along the straight line OP in Figure 7, where O is the center of subsonic domain in $[0,1] \times [0,1]$. The downward jump in the pressure is a shock front, and P is shown in Figure 2. The curve indicates that values at individual mesh points mean the shock. The crosses marked along this plot indicate pressure values at individual mesh points along the curve, in a neighborhood of the shock. They serve to show that the shock front jump is about one mesh block wide, as is typical for a numerically captured shock. In Figure 8, we represent shock strength by the difference in characteristics $\lambda_-(X) - \lambda_+(X)$ versus the angle plotted along the $M = 1$ contour to show shock existence. The angle between the ray from O and the positive x-axis is denoted by ϕ.

4.2. Generalized characteristic analysis in the 90° case. Let us recall [20]. We use the method of generalized characteristic analysis to indicate the plausible structure of the solution in the 90° case. The values imposed at infinity are given by (2.1) for the self-similar system (2.2). We first construct the solution in the far field (neighborhood of the infinity), which is comprised of four forward planar rarefaction waves R_{12}, R_{23}, R_{34}, and R_{41}, in addition to the constant states (ρ_i, u_i, v_i, p_i), $i = 1, 2, 3, 4$. We extend the four forward rarefaction waves inward from the far field till they interact, as denoted by regions in Figure 2. We find the boundary of the interaction domain, which consists of $CQRFSAEBWC$ in Figure 2, where the arc AE is an arc of sonic circle C_3 and K is the intersection point of the bottom boundaries of R_{23} and R_{41}.

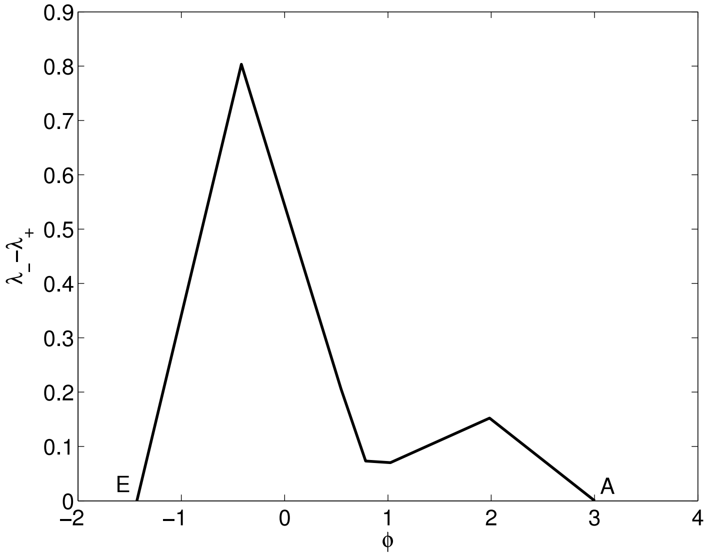


FIG. 8. *$\lambda_- - \lambda+$ vs. the angle ϕ along the $M = 1$ contour in the* 90° *case. This plot shows the shock existence; the endpoints E and A of the shock wave are as shown in Figure* 2.

Then we solve the first Goursat problem with characteristic segments CQ and CW, employing the result in [12, 15], and obtain a continuous (pseudosupersonic) solution inside the domain enclosed by the characteristic segments CQ, QD, DW, and WC. Second, we solve the Goursat problem with characteristic segments QR and QD. The solutions are still continuous in the domain $QRLD$. We continue by solving the third Gousat problem with support DL and DN. In [14], they are straight support. Then we get the continuous solutions in the domain $DLVN$. The wave R_{41} penetrates R_{12} and then R_{23} to emerge as a simple wave $RFKL$ by [14], which is adjacent to the constant state 2 and constant state 5 and located in the supersonic domain without shock wave.

We prove rigorously that two subcases possibly happen: either P_3 is greater than P_5 or P_5 is greater than P_3 in [15], where P_3 and P_5 are pressures in constant state 3 and 5. This inverted pressure profile $P_3 > P_5$ is surprising, because one would expect that pressure would be expansive in the interaction of four forward rarefaction waves. However, the inverted pressure profile is the direct result of the interaction of two waves R_{41} and R_{12}. Why is the pressure drop in the interaction region larger than the combined drop across each of the individual waves? Intuitively or based on physics, it is not easy to see whether the pressure would go up along a characteristic curve to end on a sonic point, or go down to zero to end on a vacuum. In [15], it has been proven rigorously that the pressure in the interaction region approaches zero along any characteristics, which form a hyperbolic domain determined completely by the data on the characteristic boundaries. Once the participating rarefaction waves are relatively large, the binary interaction will produce a vacuum. The pressure satisfies $P_1 > P_2 = P_4 > P_3$ and continues to drop in the simple wave interaction zone $RFKL$, which is proved in [13], to result in an even lower pressure value at K, where K is shown in Figures 2 and 6. From the numerical results, we note that the sonic boundary AP is a free boundary, as the hyperbolic domain of determinacy of the Goursat problem AFT does not include AP; see Figure 6. Thus the elliptic

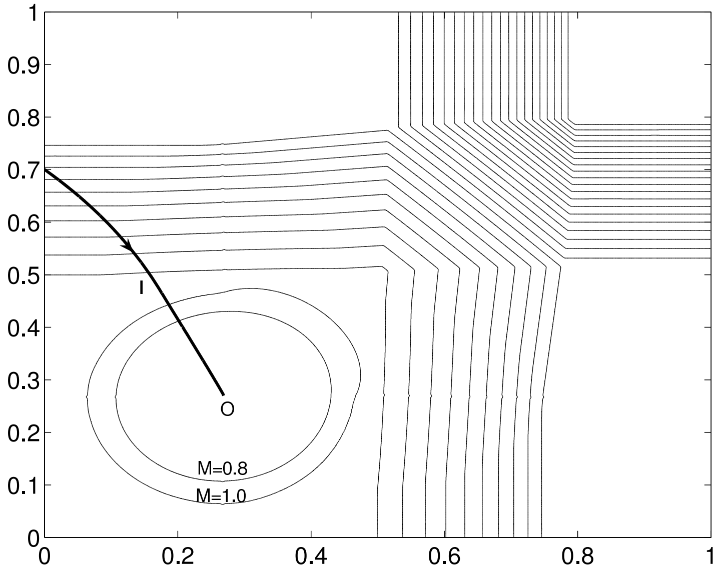


Fig. 9. *Density contours (light), two Mach contours (light) with* $M = 1.0$ *and* $M = 0.8$ *and pseudostream curve* I *(bold), which cuts through a shock wave in a neighborhood of the* $M = 1$ *contour. Arrows indicate the direction of particle motion along the stream line.*

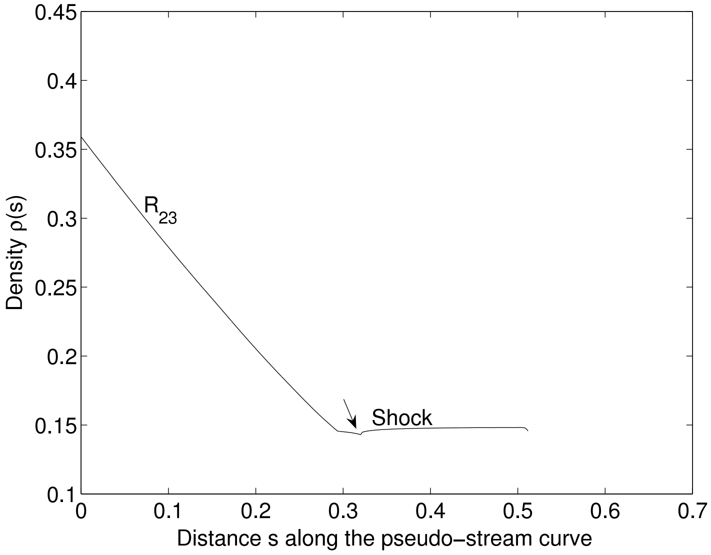


Fig. 10. *Plot of* $\rho(s)$ *vs.* s*; the position of a shock wave is visible as a small increasing bump with the distance along the pseudostream curve* I *in Figure* 9.

region influences the solution there. Numerical results show that a global minimum for the pressure in the whole space $[0, 1] \times [0, 1]$ occurs in the domain KPT. The high pressure in the subsonic domain, adjacent to the low pressure in the neighboring domain KPT and FAT, forces the shock wave to occur.

Figures 9 and 10 show the variation of density ρ along the pseudostream curve I,

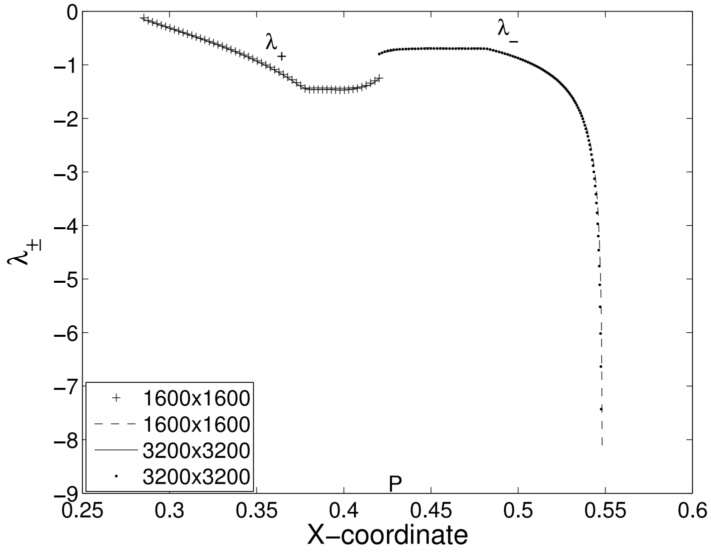


Fig. 11. *Plots of* $\lambda_+ = \frac{d\eta}{d\xi}$ *along* QP *and* $\lambda_- = \frac{d\eta}{d\xi}$ *along* PQ' *show the shock existence at* P*, since* $\lambda_+(P) \neq \lambda_-(P)$*. The plots for two computations, showing one level of mesh refinement, are indistinguishable.*

and the shock wave caused by compression waves with high pressure in the subsonic domain pushing the expansion wave in the supersonic domain. In Figure 10, s denotes the distance along the pseudostream curve.

5. Shock formation in the oblique rarefaction case.

5.1. Numerical results. In the oblique wave interaction case at $\theta = 6.5°$, two reflected characteristic curves $QP, Q'P$ in Figure 11 meet at $P|_{X=0.42}$ on the 45° diagonal line. The intersection point P is located on the $M = 1$ contour. If P were a sonic point, then we would have $\lambda_+(P) = \lambda_-(P)$. However, we show numerically in Figure 11 that P cannot be a sonic point because the characteristics are not parallel at P. At the common point P of the two plots, $\lambda_+(P) < -1$, and $\lambda_-(P) > -1, \lambda_+(P) \neq \lambda_-(P)$, indicating that P is not sonic. Thus the termination of the λ_+ characteristics and the beginning of the λ_- characteristics must be on shock. The plots for two computations, showing the level of mesh refinement, are indistinguishable. In Figure 12, we show shock strength by the difference $\lambda_-(X) - \lambda_+(X)$ in the direction of characteristics versus the angle around the $M = 1$ contour. See Figure 3 for locations of the points A, T, B, R, U, E. The plot demonstrates shock existence. The angle between the ray from O (the center of subsonic domain) and the positive x-axis is denoted by ϕ. We show further details of the nontangential termination of the characteristic curves at the shock front in Figure 13. The characteristic curves terminate nontangentially and are not parallel to each other. We find numerically that the reflected simple wave is a compressive wave and forms a weak shock. See Figure 14, where the characteristic distance denotes the "shock distance." The separation distance, i.e., the normal separation between two neighboring characteristics, is plotted versus the length along the reflected characteristics. The plot also shows the occurrence of the shock.

We plot pressure p against the distance from the origin $R = (x^2 + y^2)^{\frac{1}{2}}$ along the 45° diagonal line in Figures 15 and 16. Under refinement of the mesh, the oscillations

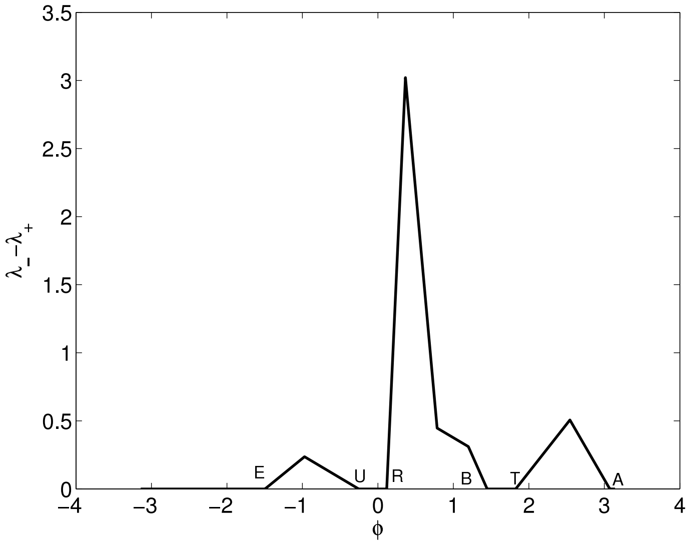


FIG. 12. *The difference* $\lambda_-(X) - \lambda_+(X)$ *vs. angle along the* $M = 1$ *contour at* $\theta = 6.5°$. *This plot shows existence of shocks both on the inward and the reverse sides of the* $M = 1$ *contour, in the second and fourth quadrants.*

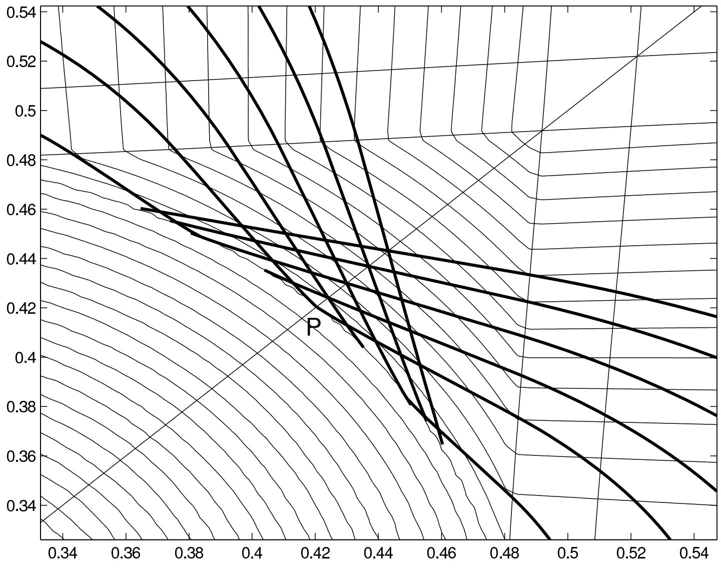


FIG. 13. *Enlarged view of Figure* 3 *with details near the point* P *on the shock front at* $6.5°$. *Bold curves are* $\lambda_\pm$ *characteristics; light curves are Mach number contours. Note that the characteristics terminate nontangentially on the shock.*

get weaker and the shock becomes sharper. The circles and crosses are located at mesh block centers, for cells within the shock profile. The trend of convergence of shocks in each plot in Figure 15 is clear and sufficient: the shock wave here is very weak, but its strength is not decreasing as the mesh is refined; the shock will be stable even with extremely fine meshes.

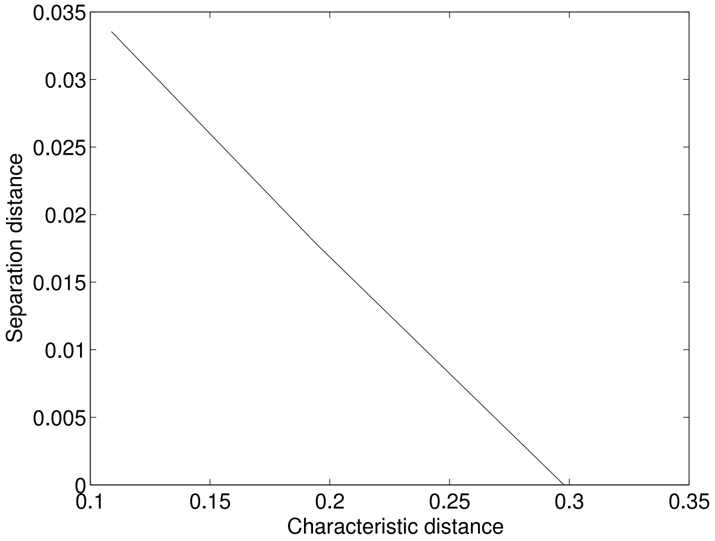


FIG. 14. *Plot of separation between neighboring characteristics vs. distance along characteristics with $\theta = 6.5^\circ$ in case B. This plot shows shock formation.*

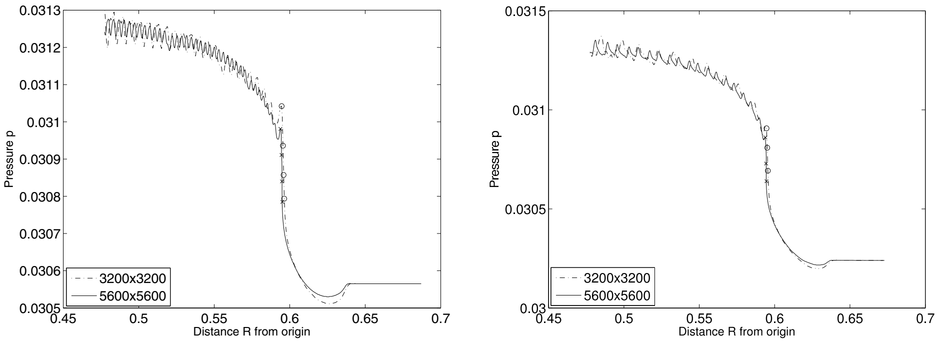


FIG. 15. *Left: pressure vs. distance along a* 45° *diagonal line at* 6°*. Right: pressure vs. distance along a* 45° *diagonal line at $\theta = 6.5^\circ$. The x and o indicate cell center solution values moving through the shock for the region of rapid solution transition.*

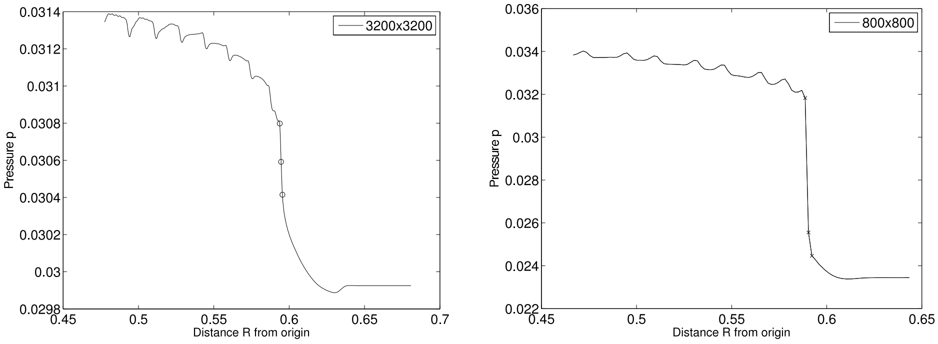


FIG. 16. *Pressure vs. distance along a* 45° *diagonal line. Left: $\theta = 7^\circ$. Right: $\theta = 22.5^\circ$.*

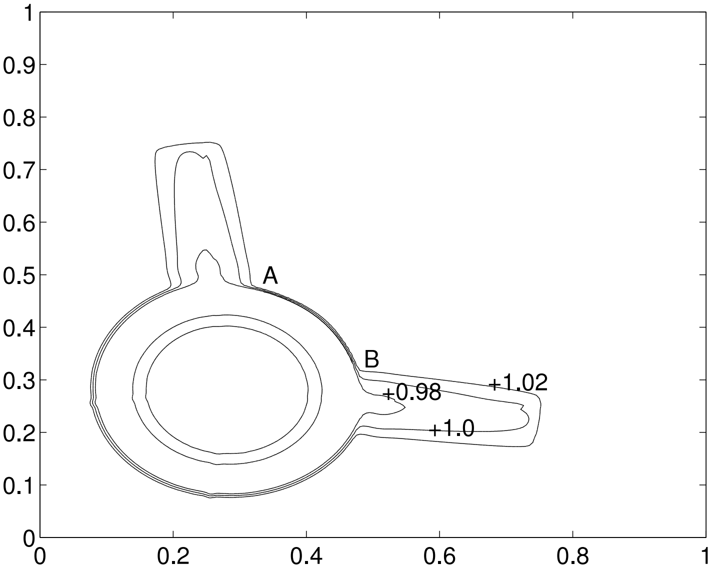


Fig. 17. *Comparison of wave filter shock location A, B and pseudo-Mach number contour plots at $\theta = 8^\circ$ with an 800×800 mesh.*

We use the wave filter embedded in the FronTier code. The wave filter is an automated pattern recognition algorithm which locates shock waves, rarefaction waves, and contact discontinuities in numerical solutions of the Euler equations for compressible fluids on the basis of detecting a local jump in the solution which satisfies the Rankine–Hugoniot relations. The shock wave as determined by this wave filter program is shown in Figure 17 by the curve AB. Note that the labeled Mach number contours $M = 0.98$ and $M = 1.02$ in Figure 17 and $M = 0.96$ and $M = 1.02$ in Figure 18 coincide on the curve AB, indicating that they are shock fronts. These curves match the pseudo-Mach number contours well.

5.2. Generalized characteristic analysis. We use the method of generalized characteristic analysis to indicate the plausible structure of the solution to our problem for $\theta > 8^\circ$ based on numerical results in section 5.1. We retain the notation from [20]. We discuss causal relationships and decompose the boundary value problem into three subproblems based on the features of the characteristics. The analysis consists of the following steps. The first is a classical rarefaction Goursat problem which has been solved analytically in [12, 15]. The second is a degenerate Goursat problem, whose solution is proved to be a simple wave in [14]. The last is a pseudotransonic boundary value problem with free boundaries consisting of interior sonic curves and shocks. The mathematical proof for the structure of this last subproblem is open. The problem involves collisions of rarefaction waves with sonic curves which produce compressive waves upon reflection, which may then form shocks. We outline the boundary of the domain of interaction for the initial four rarefaction waves in both cases above.

Step 5.2.1. *The constant states and simple waves in the far field.* As in [20], we transform problem (2.1) and (2.3) into a boundary value problem for the self-similar system (2.2) with values imposed at infinity,

$$\lim_{\xi^2+\eta^2\to\infty} (\rho, u, v, p)(\xi, \eta) = (\rho_i, u_i, v_i, p_i), \quad i = 1, 2, 3, 4, \tag{5.1}$$

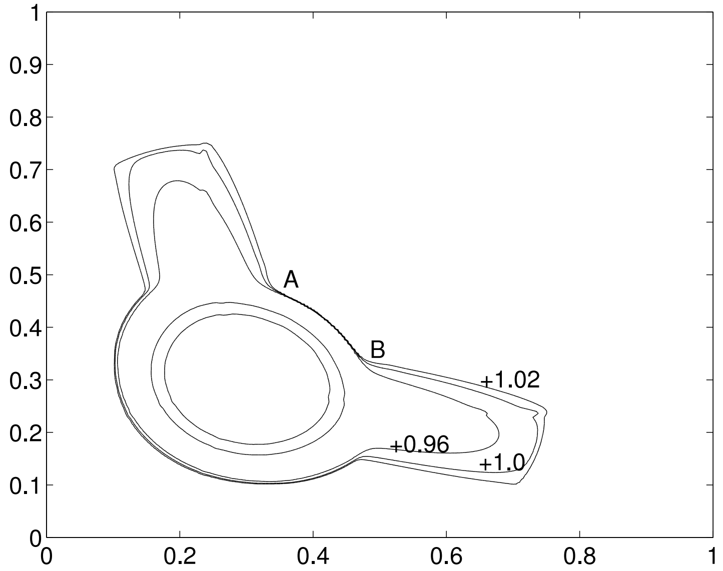


FIG. 18. *Comparison of wave filter shock location A, B and pseudo-Mach number contour plots at $\theta = 22.5^\circ$.*

in which the limiting direction is consistent with the data sector in (2.3). We first construct the solution in the far field (neighborhood of the infinity), which is comprised of four forward planar rarefaction waves R_{12}, R_{23}, R_{34}, and R_{41}, besides the constant states (ρ_i, u_i, v_i, p_i), $i = 1, 2, 3, 4$. We extend the four forward rarefaction waves inward from the far field till they interact, denoted by regions as in Figure 19. We find the boundary of the interaction domain, as in [20], which consists of the characteristic segments PQ, QR, ST, TU, $U'T'$, $T'S'$, $R'Q'$, $Q'P$ and arcs of sonic circles RS, UU', $S'R'$. See Figure 19.

Step 5.2.2. *Simple wave solutions after interaction of planar rarefaction waves.* The two rarefaction waves R_{12} and R_{41} start to interact at P. We use the result in [12, 15] to solve the Goursat problem with the boundary values supported on the characteristic curves PQ and PQ', and obtain a continuous (pseudosupersonic) solution inside the domain enclosed by the characteristic segments PQ, QP', $P'Q'$, and $Q'P$.

Then we proceed to solve the Goursat problem with the boundary data supported on QR and QP'. Since the state (ρ_2, u_2, v_2, p_2) is constant, we use the result in [14, Theorem 7]: *Adjacent to a constant state is a simple wave in which (ρ, u, v, p) are constant along a family of wave characteristics which are thus straight.* This fact indicates that the solution is a simple wave, denoted by R_{25}, in the angular domain between QR and QP'. We note that the simple wave R_{25} just covers the region of the curvilinear quadrilateral $QRWP'$ from the theory of characteristics, where RW is the λ_+-characteristic curve from the point R and can be regarded as the reflection of the λ_--characteristic curve at that point. Also note that the point R on the sonic curve C_2 is degenerate. It has the following interesting properties: It is of Tricomi type from the side of R_{25}, but of Keldysh type from the side of the constant state (ρ_2, u_2, v_2, p_2). A point on a sonic curve is said to have a Tricomi type if the characteristics are nontangential to the sonic curve. It is called Keldysh type if the characteristics are

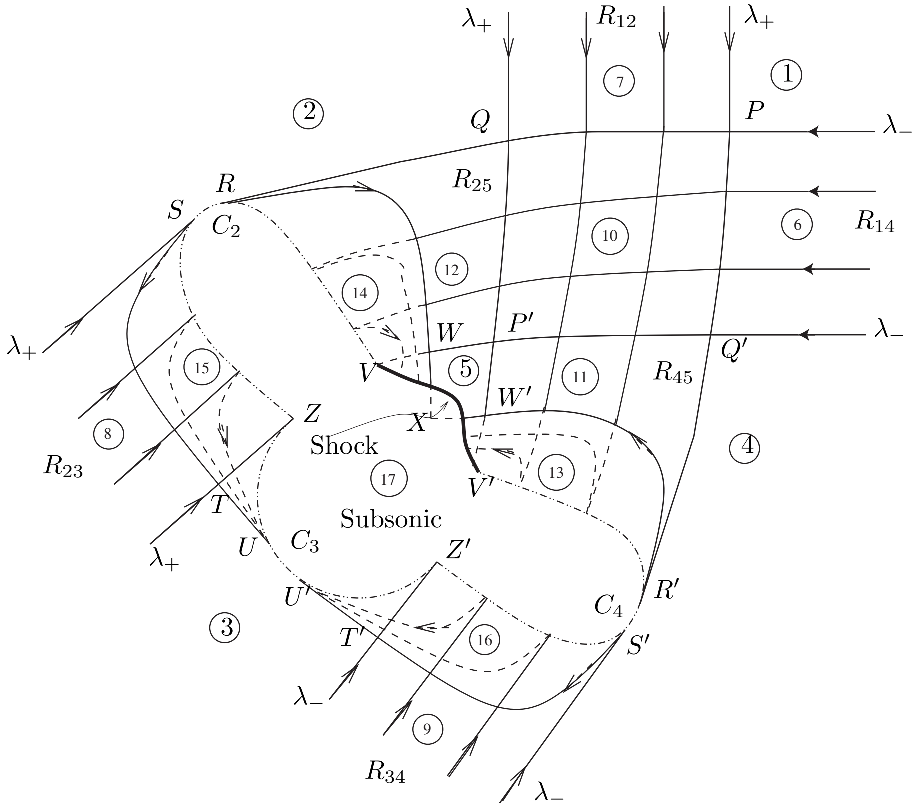


FIG. 19. *Generalized characteristic analysis for the case of four forward rarefactions in a* 2D *Riemann problem.*

tangential to the sonic curve.

We continue to solve the Goursat problem with the support of two straight characteristic curves $P'W$ and $P'W'$. Obviously, the solution is a constant state (ρ_5, u_5, v_5, p_5) with boundary $P'WXW'$. The point X must be outside the sonic circle of the state (ρ_5, u_5, v_5, p_5).

Step 5.2.3. *Plausible solution structure in intersecting supersonic regions with the transonic boundary.* After the above three Goursat problems, we reach the boundary $RWXW'R'$.

Now we consider the problem of the pseudotransonic flow with the boundary $RSTUU'T'S'R'W'XWR$. It is reasonable to assume a priori that the family of the λ_--characteristic curves of the simple wave R_{25} extends to the sonic curve RV and reflect off it as a family of λ_+-characteristic curves. Here the extension of R_{25} is not a simple wave because the solution must vary along both λ_- and λ_+ characteristics. The solution is jointly determined by the subsonic domain and the supersonic domain. These reflected λ_+-characteristic curves reach a curved boundary WV, which forms another degenerate Goursat problem with the support of a straight λ_+-characteristic line $WX$and a curved λ_--characteristic line WV. This degenerate Goursat problem has a simple wave solution whose data are on the left boundary QP' since adjacent to the WX side is the constant state (ρ_5, u_5, v_5, p_5). The numerical results indicate that this reflected simple wave is a compressive wave and forms a shock with starting

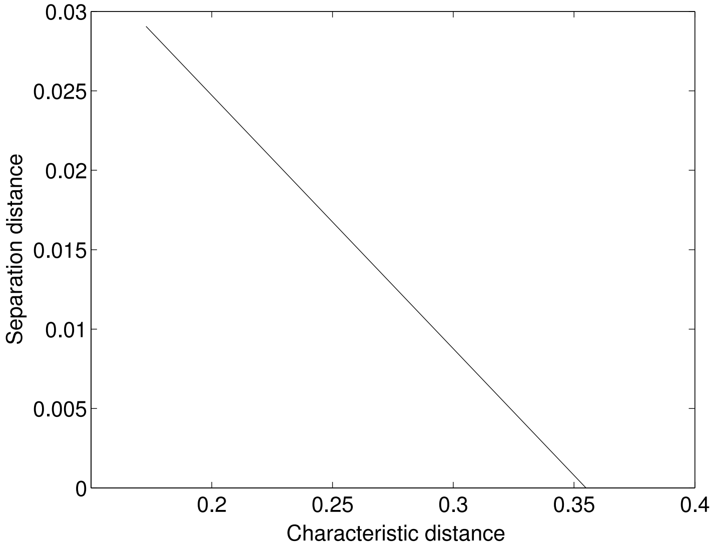


FIG. 20. *Plot of separation distance between neighboring characteristics starting on the sonic curve vs. distance along characteristics with $\theta = 22.5°$.*

point V. See Figure 20, where characteristic distance denotes the shock distance, i.e., the length of the reflected characteristics, and separation distance is the normal separation between two neighboring characteristics. This plot shows the occurrence of the shock. The shock borders the constant domain (ρ_5, u_5, v_5, p_5). By symmetry, this structure is repeated across the symmetric axis with starting point V' of another shock in the primed variables $W'V'X$.

We analyze the numerical results in Figures 21 and 22, which show the variation of density ρ along pseudostream curves I and II, presenting regions corresponding to expansion and compression waves. The arrows on the stream curves indicate the direction of particle motions. In Figure 22, the distance s denotes the distance along pseudostream curves, standing at the data beginning and ending on the pseudostream curves I and II.

The structure of the solution for the reflected characteristic curves R_{23} in Figure 19 is clarified, where the family of λ_+-characteristic curves coming from R_{23} collide with a Tricomi-type pseudosonic curve SZ and are reflected to form a weak shock wave ZU, which resembles the 90° case.

6. Shock formation for two backward and two forward rarefaction waves.

For the case of two backward and two forward rarefaction waves, there are two symmetric transonic shocks in the solution, as shown in [1, 2, 17, 11, 13]; see Figure 23. The mechanism of shock formation is the same as was discussed for the case of four forward rarefaction waves in Figure 19, because the part of Figure 23 upper-right to $\xi + \eta = u_2 + v_2$ has the same structure as that of the corresponding part of Figure 19.

7. Conclusion.

We discuss numerical simulations showing two distinct mechanisms for shock formation and supporting theoretical conjectures based on generalized characteristic analysis regarding mathematical mechanism in the 90° case and the oblique rarefaction case. We also discover that the same mathematical mechanism as in the oblique rarefaction case occurs for the shock formation for two forward and two backward rarefaction waves.

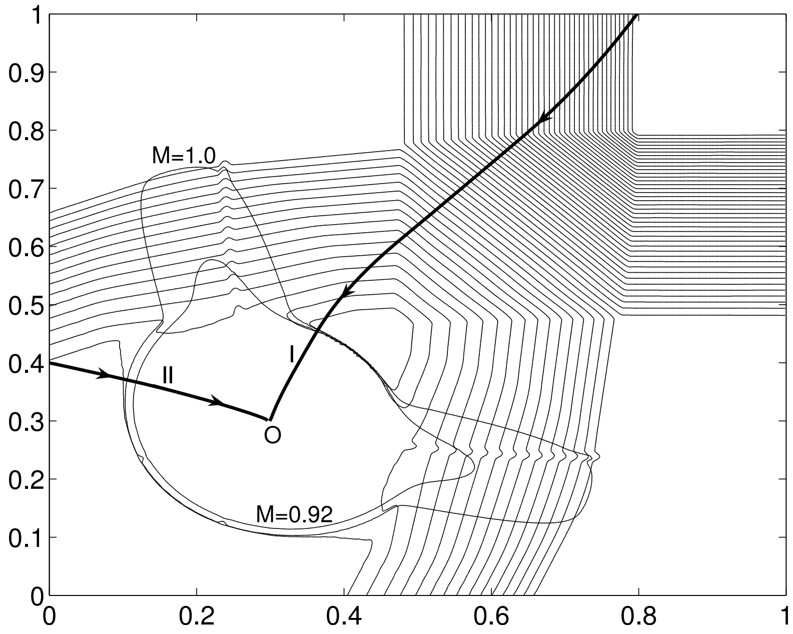


Fig. 21. *Density contours (light); two Mach contours (light) with $M = 1$, $M = 0.92$; and two pseudostream curves I, II (bold), which cut through the weak shock waves and shock waves in the neighborhood of the $M = 1$ contour shown at $\theta = 22.5°$.*

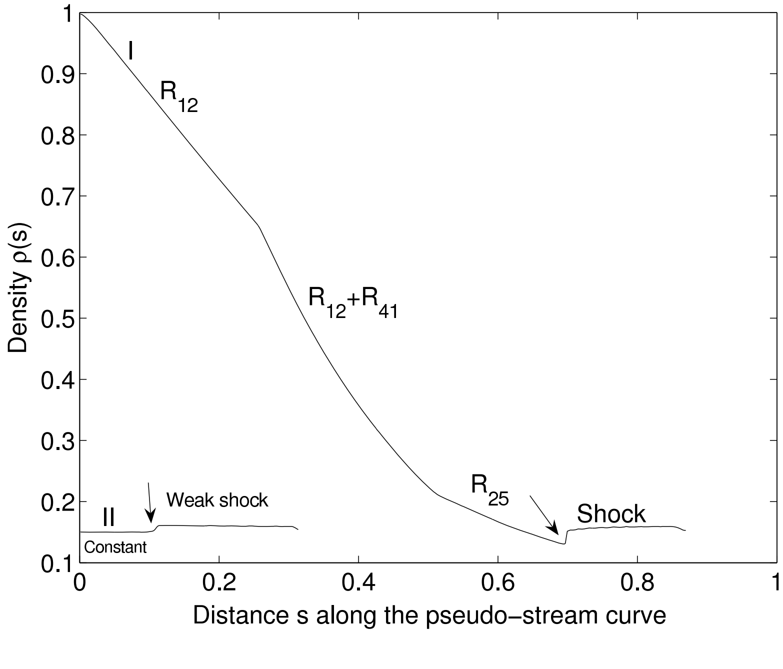


Fig. 22. *The increasing positions of (weak shock wave and shock wave) in the $\rho(s)$ vs. s plots are visible as bumps along two pseudostream curves I (above) and II (below) in Figure 21.*

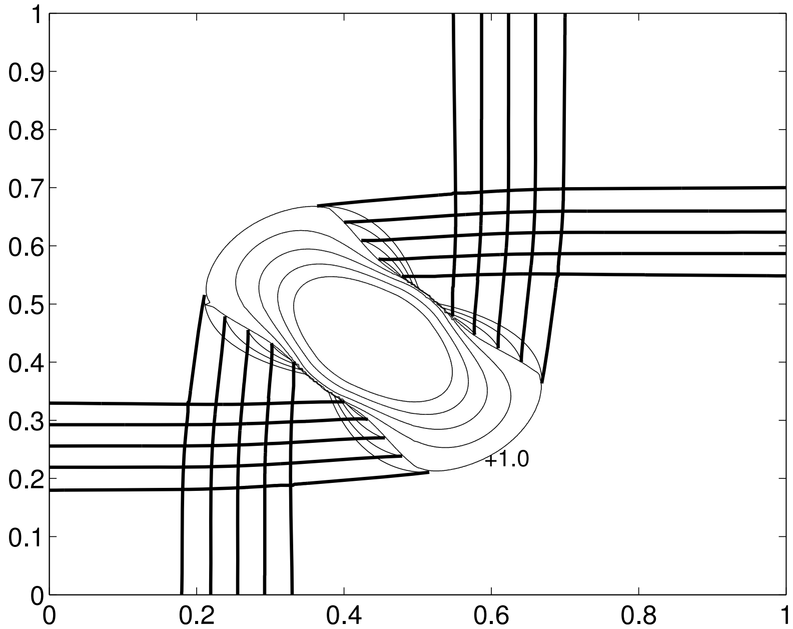


FIG. 23. *The case of two forward and two backward rarefactions oriented at* 90° *with* $p_1 = 0.444, \rho_1 = 1.0$*,* $u_1 = v_1 = 0.00$*,* $\rho_2 = 0.5197$*,* $T = 0.25$*; characteristics (bold) and contour curves of pseudo-Mach number (light) are plotted.*

For the 90° case, the interaction of two rarefaction waves of the same family and parallel at infinity leads to a pressure drop larger than that due to enter taken singly. Thus the interaction seems to "overrarefy," leading to low pressure states incompatible with pressures given at infinity due to the same rarefactions considered individually. A shock wave results from the joining of these high and lower pressure regions. It is the interaction of rarefaction R_{41}, R_{23}, R_{34}, and R_{12}, including the interaction of R_{41} and characteristics from constant state 3 below R_{23}, which produces this result.

The shock formation for the oblique case has a possible physical mechanism similar to one found in stationary flow, which is illustrated schematically in Figure 24, associated with the numerical result in Figure 4. Basically, a rarefaction reflection reflects at a sonic boundary; the reflected wave is a compression, which may in time break and become a shock. For the steady transonic small disturbance equation, shock reflection on a sonic curve is illustrated in Cole and Cook [4]; see the shock formation over an airfoil in Figure 5.4.13, p. 314 there. The structure of shock formation from the reflection of rarefaction waves on a sonic curve was suggested by Guderly [9]; for the 2D steady irrotational isentropic flow, they put forward a concept of shock formation from reflection of characteristics on a sonic curve. When a supersonic bubble appears on the top of an airfoil in an ambient subsonic domain, a family of characteristics are generated in the bubble from the surface of the airfoil, and they hit the rear portion of the sonic curve and are reflected downstream to form a compressive wave, which then forms a shock wave within the bubble. The subsonic region plays the role of a permeable obstacle which declines streamlines towards the airfoil and causes compressive waves. This is similar to the way a concave wall does in the classical problem of supersonic flow over a smooth rigid wall. We point out

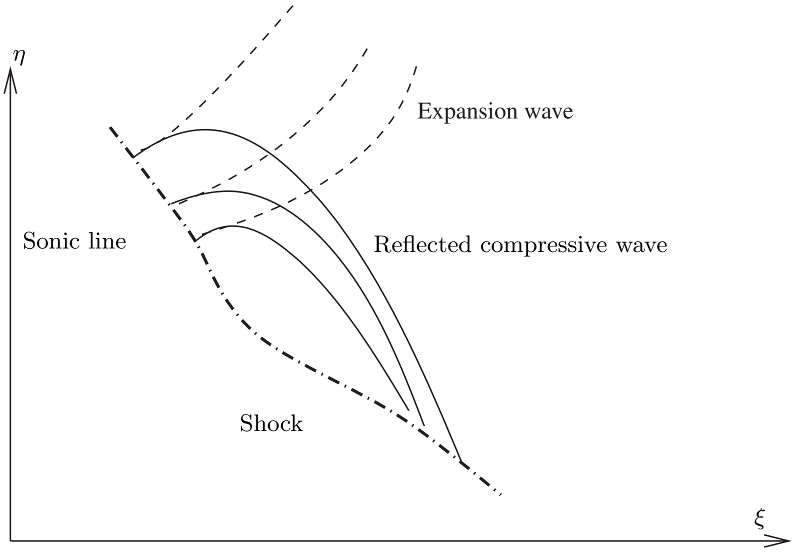


FIG. 24. *Shock formation from the reflection of rarefaction wave on a sonic curve.*

that the bump on the wall of the flow channel causes the shock formation naturally in Cole and Cook [4], where it has an important application in the study of a flow over an airplane wing. Furthermore, this formation of shocks seems to be the fundamental mechanism for the Guderly reflection pattern in Tesdall and Hunter [18].

For waves interacting with a sufficiently large oblique angle in the third quadrant, the sonic curve has an exaggerated nonconvex shape (rabbit ears). The curve extends into the rarefaction waves and interacts with them. The rarefactions are reflected as compression waves along these rabbit ears, in the sense that along this part of the sonic curve there are impinging λ_+ and λ_- characteristics. The characteristics coming from infinity are part of the rarefaction wave, while the reflected ones, as stated, are compressions. On the side facing the first quadrant, these compressions have sufficient travel distance to break and form a shock wave, centered at the 45° line, where it crosses the $M = 1$ contour. On the side facing the third quadrant, due to the angle between the waves at infinity, there is a very weak shock on this side of the $M = 1$ contour, which can be analogously interpreted in Figure 11 of [3, p. 390], where stationary flow is in nozzles and jets. Very likely these characteristics would have an envelope if they were not intercepted by a shock front. To prevent the envelope singularity, an "intercepting" shock is therefore necessary. The reason for the reflected wave to be a compression is illustrated by similarity to a related problem in aerodynamics, as discussed in the literature and explained in [4] and [9].

The Riemann problem is typically unstable in that it is a locus of bifurcation for the Riemann data. Even in one dimension, the isolated jump discontinuity holds only at time zero and (for gas dynamics) the solution at all positive times has three traveling waves.

However, it is stable in the sense of preservation of structure upon variation of initial (Riemann) conditions. In this sense, our analysis deals with representative variation of the initial conditions but does not explore the complete 7D space of nearby initial conditions numerically. This numerical stability analysis was conducted for the full Euler equations (2.1) rather than for the self-similar equations (2.2). No non–

self-similar solutions were observed in the variations of initial data. We have studied systematically a variation of the angle between two of the four initial rarefaction waves. As this angle is modified sufficiently, we find a jump to a new solution branch, with a distinct mechanism (in a detailed sense) for the shock formation. Although it is conceptually possible to allow an arbitrary number of waves, at arbitrary angles in the 2D Riemann problem formulation, the case of four waves is the case most commonly considered.

Acknowledgment. Discussions with John Hunter and Wancheng Sheng were very helpful.

REFERENCES

[1] T. Chang, G. Chen, and S. Yang, *On the 2-D Riemann problem for the compressible Euler equations,* I. *Interaction of shocks and rarefaction waves*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 555–584.

[2] T. Chang, G. Chen, and S. Yang, *On the 2-D Riemann problem for the compressible Euler equations,* II. *Interaction of contact discontinuities*, Discrete Contin. Dynam. Systems, 6 (2000), pp. 419–430.

[3] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer-Verlag, New York, 1948.

[4] J. D. Cole and L. P. Cook, *Transonic Aerodynamics*, North–Holland, Amsterdam, 1986.

[5] P. Colella, *A direct Eulerian MUSCL scheme for gas dynamics*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 104–117.

[6] J. Glimm, J. W. Grove, Y. Kang, T. Lee, X. Li, D. H. Sharp, Y. Yu, K. Ye, and M. Zhao, *Statistical Riemann problems and a composition law for errors in numerical solutions of shock physics problems*, SIAM J. Sci. Comput., 26 (2004), pp. 666–697.

[7] J. Glimm, H. C. Kranzer, D. Tan, and F. M. Tangerman, *Wave fronts for Hamilton–Jacobi equations: The general theory for Riemann solutions in R^n*, Comm. Math. Phys., 187 (1997), pp. 647–677.

[8] J. Glimm and D. Sharp, *An S-matrix theory for classical nonlinear physics*, Found. Phys., 16 (1986), pp. 125–141.

[9] K. G. Guderly, *The Theory of Transonic Flow*, Pergamon Press, London, 1962.

[10] A. Kurganov and E. Tadmor, *Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 548–608.

[11] P. D. Lax and X.-D. Liu, *Solution of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM J. Sci. Comput., 19 (1998), pp. 319–340.

[12] J. Li, *On the two-dimensional gas expansion for compressible Euler equations*, SIAM J. Appl. Math., 62 (2001), pp. 831–852.

[13] J. Li, T. Zhang, and S. Yang, *The Two-Dimensional Riemann Problem in Gas Dynamics*, Pitman Monographs 98, Longman Press, London, 1998.

[14] J. Li, T. Zhang, and Y. Zheng, *Simple waves and a characteristic decomposition of the two dimensional compressible Euler equations*, Comm. Math. Phys., 267 (2006), pp. 1–12.

[15] J. Li and Y. Zheng, *Interaction of rarefaction waves of the two-dimensional self-similar Euler equations*, Arch. Ration. Mech. Anal., to appear.

[16] X. Liu and P. D. Lax, *Positive schemes for solving multi-dimensional hyperbolic systems of conservation laws*, J. Comp. Fluid Dynam., 5 (1996), pp. 133–156.

[17] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., 14 (1993), pp. 1394–1414.

[18] A. M. Tesdall and J. K. Hunter, *Self-similar solutions for weak shock reflection*, SIAM J. Appl. Math., 63 (2002), pp. 42–61.

[19] Y. Yu, M. Zhao, T. Lee, N. Pestieau, W. Bo, J. Glimm, and J. W. Grove, *Uncertainty quantification for chaotic computational fluid dynamics*, J. Comput. Phys., 217 (2006), pp. 200–216.

[20] T. Zhang and Y. Zheng, *Conjecture on the structure of solutions of the Riemann problem for two-dimensional gas dynamics systems*, SIAM J. Math. Anal., 21 (1990), pp. 593–630.

[21] Y. Zheng, *Systems of Conservation Laws: Two-Dimensional Riemann Problems*, Progress in Nonlinear Differential Equations and Their Applications 38, Birkhäuser Boston, Cambridge, MA, 2001.

SIAM J. APPL. MATH.
Vol. 69, No. 3, pp. 743–762

FILTERS. THE NUMBER OF CHANNELS THAT CAN CLOG IN A NETWORK*

GUIDO KAMPEL† AND GUILLERMO H. GOLDSZTEIN†

Abstract. We model filters as two-dimensional networks of channels. As a suspension (fluid with particles) flows through the filter, particles clog channels. We assume that there is no flow through clogged channels. In this paper, we compute a sharp upper bound on the number of channels that can clog before fluid can no longer flow through the filter.

Key words. filters, porous media, networks, clogging, planar graphs, Euler formula

AMS subject classification. 76S05

DOI. 10.1137/080723703

1. Introduction.

A porous medium is a material that contains relatively small spaces filled with fluid embedded in a solid matrix. These fluid-filled spaces are called pores. A porous material is said to be permeable if fluid can flow through its pores from one end to an opposite end of the material. Filters are examples of porous materials.

Fluid suspensions (or suspensions, for short) are fluids with small solid particles in them. According to their size and properties, these particles are called fines or colloids. As a suspension flows through a permeable porous material, some fines are trapped within the material. In fact, the function of the filters we consider in this paper is to *clean* suspensions by capturing most particles bigger than a certain size.

The removal of particles from fluid suspensions is of importance in a wide range of industrial and technological applications such as waste water treatment [18] and other filtration processes [4, 31]. Our studies are motivated by the filters used in the process known as deep bed filtration. As a suspension flows through a filter composed of granular or fibrous materials, fines or colloidal particles penetrate the filter and deposit at various depths [32]. As a result, the fluid suspension is cleaner when it exits the filter (i.e., it exits the filter with many fewer solid particles than it originally had when it entered the filter).

Theoretical models to study transport in porous media can be classified as either macro-scale [5, 15, 22, 23, 24, 26, 32] or pore-scale [9, 19, 28] models. Within the latter group, the class of network models, in which the pore space is modeled as a network of channels, is very popular. Network models provide flexibility in modeling different geometries of pore space while keeping the computational cost low. Our work belongs to this class of models.

Network models to study transport in porous media were introduced by Fatt in 1956 [10, 11, 12]. Donaldson, Baker, and Carrol in 1977 [8] were the first to use networks to study particle transport within porous media. The clogging of particles has been studied in networks with different geometries including bundles of parallel tubes [8], square networks [14, 16, 21], triangular networks [3, 25], cubic networks

*Received by the editors May 8, 2008; accepted for publication (in revised form) August 26, 2008; published electronically December 17, 2008. This research was supported by the NSF.

http://www.siam.org/journals/siap/69-3/72370.html

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (kampel@math.gatech.edu, ggold@math.gatech.edu).

[2, 17, 29], bubble models [6, 20], and the so-called three-dimensional physically representative networks [1, 30].

Consider a filter that is a network of channels. As a suspension flows through the filter, particles clog channels. Assume that the suspension cannot flow through clogged channels. In this work, channels that are not clogged are called open. Note that there can be flow only through channels that are part of a percolating path of open channels, i.e., a path of channels that are not clogged connecting one side of the filter with the opposite side. As channels clog, some percolating paths of open channels are broken. Thus, suspension stops flowing not only through the clogged channels, but also through other channels, i.e., those that are no longer part of a percolating path of open channels. Thus, the filter will stop being permeable after not all, but only a number, of its channels clog. In this paper we find an upper bound of this number. Our upper bound is a function of the geometry of the network. In particular, we are able to identify the filter geometries for which the largest fraction of channels may be clogged before the filter ceases to be permeable. Our results suggest that filters with these geometries may have longer lives than others.

Our work is novel. Most of the work that can be found in the literature consists of simulations of the suspension dynamics within the medium. Our work is an analysis that is independent of the dynamics; it depends only on the topology of the network. On the other hand, our work has connections to, but also key differences from, the theory of bond percolation [13, 27]. In percolation theory, channels or edges are removed randomly and independently of each other. Here, channels clog, but neither randomly nor independently of each other; the order in which they clog is important. Nevertheless, we are able to use graph theory techniques that are also used in percolation theory.

We remark that in this paper we assume the porous media to be two-dimensional. Extensions to three-dimensional media, which could lead to results more relevant to real applications, are currently being pursued and will be presented elsewhere. We also acknowledge that our work ignores the dynamics and does not resolve the mechanisms of clogging, which would involve a variable flow-field, drag forces, and particle-solid interaction forces that are all part of a well-developed filtration theory.

This paper is organized as follows. In section 2, we describe the filters as networks. In section 3, we review the basics of graph theory that are needed in the rest of this paper. In section 4, we obtain our upper bound. In section 5, we show that our upper bound is sharp. In section 6, we consider large filters and obtain an alternative description of our bound in terms of the average degree of the network. In section 7, we consider a special class of filters for which our bound is realized. In section 7, we also consider some examples and obtain some conclusions.

2. The model.

We model filters as two-dimensional networks of channels as we illustrate in Figure 1. The pores are the interiors of the channels. Our filters have a bottom boundary at $y = y_b$ and a top boundary at $y = y_t$.

In our model, channels are either open or clogged. Suspension can flow only through open channels. There is no flow through clogged channels. Within an open channel, suspension flows from the end with higher pressure to the opposite end. If both ends are at the same pressure, there is no flow within the channel.

We assume that suspension can flow into the filter only through the bottom boundary and can flow out of the filter only through the top boundary. Both fluid and particles are incompressible, and thus a volume of suspension enters the filter through the bottom boundary at the same rate it exits the filter through the top boundary.

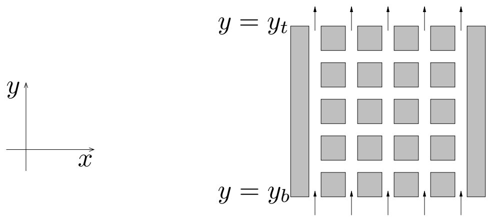


Fig. 1. *As illustrated in the left figure, we use the standard notation of x-axis and y-axis for the horizontal and vertical axes, respectively. The right figure shows a network of channels. The arrows indicate the direction of the flow.*

We assume that the bottom boundary is held at constant pressure $p = p_b$ and the top boundary at $p = p_t$, where $p_b > p_t$. Note that the filter is permeable if and only if there is a path of open channels connecting the bottom boundary with the top boundary. Due to the difference in pressure between the top and bottom boundaries, there is flow through the filter if and only if the filter is permeable.

We assume that initially all the channels are open. As suspension flows through the filter, particles are trapped, causing channels to clog; i.e., channels change from open to clogged. Eventually, the filter is no longer permeable. Note that an open channel can clog only if there is flow through it. For any given filter, we will find an upper bound on the number of channels that may clog under the assumption that different channels do not clog simultaneously.

Assumptions 2.1. *For future reference, we list here the key assumptions of our model:*

1. *Channels are either open or clogged.*
2. *There is no flow through clogged channels.*
3. *Suspension can only flow into the filter through the bottom boundary and out of the filter through the top boundary.*
4. *Fluid and particles are incompressible.*
5. *Initially all the channels are open.*
6. *An open channel may clog if there is flow through it.*
7. *An open channel does not clog if there is no flow through it.*
8. *Different channels do not clog simultaneously.*

3. Review of concepts in graph theory.

In this section we review concepts of graph theory that we need in the rest of the paper. More details on graph theory can be found in [7].

A *graph G* consists of a nonempty set of elements, called *vertices* or *nodes*, and a list of unordered pairs of these elements, called *edges*. It is convenient and a common practice to draw graphs in the plane. Each node is a different point in the plane, and each edge a line joining its two nodes without intersecting any other node. If e is an edge joining the two nodes a and b, we say that a and b are the end points of e and that e connects a and b. For convenience we take e (the drawing of e really) to be a closed set; i.e., e includes its end points. If $a = b$, i.e., the end points of an edge e are the same, we say that e is a loop. In a graph, two different edges do not have the same pair of end points. We have a *multigraph* when this restriction is removed; i.e., in a multigraph, two different edges can have the same end points.

We say that two nodes a and b are connected if there exists a sequence of nodes $n_0, n_1, \ldots, n_k$ such that $a = n_0$, $b = n_k$ and for each $1 \leq i \leq k$ there exists an edge e_i

that connects n_{i-1} and n_i. In this case, the alternating sequence of nodes and edges $n_0, e_1, n_1, e_2, n_2, \dots, e_k, n_k$ forms a *walk* between a and b or simply a *walk*. We say that $a = n_0$ and $b = n_k$ are the end points of the walk. If $n_i \neq n_j$ for all $i \neq j$, we say that the walk is a *path*. If $n_0 = n_k$ and $n_i \neq n_j$ for $i < j$ except when $(i, j) = (0, k)$, we say that the walk is a *cycle*. We will identify each walk with the curve in the plane formed by its edges.

Let G be a multigraph. S is a *submultigraph* of G if S is a multigraph and S is included in G, i.e., every node of S is also a node of G and every edge of S is also an edge of G.

A multigraph is *connected* if there is a walk between any pair of its nodes, and *disconnected* otherwise. Every multigraph is the disjoint union of connected submultigraphs. Each of these submultigraphs is called a *connected component* of the multigraph.

A multigraph is *planar* if it can be drawn in the plane in such a way that any two different edges may intersect at only one or two of their end points. Any such drawing is a *plane drawing* of the multigraph. In this paper we will need to consider only planar multigraphs. We identify each planar multigraph with one of its plane drawings. In the rest of this paper, any multigraph that we mention or consider is a planar multigraph.

A multigraph divides the plane into regions called *faces*. More precisely, the faces are the connected components of what is left from the plane once we remove the multigraph from the plane. In other words, the faces are the connected components of the set of points in the plane that do not belong to any edge of the multigraph. Note that the faces are open sets. Any finite multigraph has an unbounded face surrounding it, called the *infinity face*.

Note that the boundary of any bounded face contains a cycle. Thus, a connected multigraph with no cycles has only one face, the infinity face.

Let G be a multigraph. We denote by n_G its number of nodes, by e_G its number of edges, by f_G its number of faces, and by ℓ_G its number of connected components. The well-known *Euler formula* states that

$$n_G + f_G = e_G + \ell_G + 1. \tag{3.1}$$

The degree of a node n, which we denote by d_n, is the number of edges that have n as an end point, where the loops are counted twice. The average degree of a multigraph G, which we denote by d_G, is defined as the average of the degrees of the nodes of G, $d_G = n_G^{-1} \sum d_n$, where the sum is over all nodes n and n_G is the number of nodes of G. Note that

$$d_G = 2\frac{e_G}{n_G}, \tag{3.2}$$

where e_G is the number of edges of G. An example of a multigraph that is actually a graph is shown in Figure 2.

4. Upper bound on the number of clogged channels.

4.1. Microstructure of the filters. To each filter we associate a multigraph in a natural way. The edges are the channels and the nodes the end points of the edges.

Recall that the bottom and top boundaries of the filter are located at $y = y_b$ and $y = y_t$, respectively. Thus, the multigraph is included in $y_b \leq y \leq y_t$. Note that there

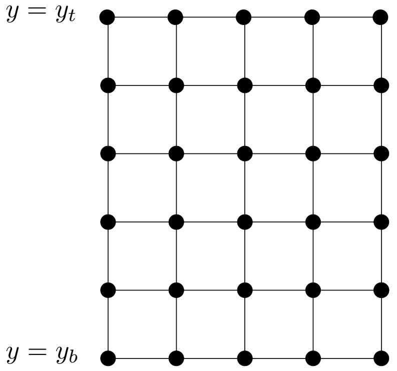


FIG. 2. *Multigraph G. The black small circles are the nodes of the multigraph G and the solid lines its edges.*

are nodes in the bottom and top boundaries. For convenience, we also include edges in $y = y_b$ connecting the nodes in the bottom boundary. In other words, there is a path of edges in $y = y_b$ connecting the leftmost node in the bottom boundary with the rightmost node in that boundary. Analogously, we include edges in $y = y_t$ so that there is a path of edges in $y = y_t$ connecting the leftmost node in the top boundary with the rightmost node in that boundary.

We consider filters with a finite number of channels, and thus our multigraphs are finite multigraphs; i.e., they contain a finite number of nodes and edges. As an example, in Figure 2 we show the multigraph G associated with the filter of Figure 1.

DEFINITION 4.1. *We say that a node is an exterior node if it is located at $y = y_b$ or $y = y_t$. Otherwise, we say that the node is an interior node.*

We also say that an edge is an exterior edge if it is included in $\{y = y_b\} \cup \{y = y_t\}$. Otherwise, we say that the edge is an interior edge.

Note that, by construction, each exterior node is the end point of at least one interior edge.

4.2. Clogged edges. The suspension enters the network through exterior nodes at $y = y_b$ and exits the network through exterior nodes at $y = y_t$.

We say that an edge is clogged if the corresponding channel is clogged. Note that the exterior edges were included for convenience; they do not correspond to any channel. Thus, we assume that there is no flow within them and that they never clog; i.e., they are always open.

As we said in section 2, we assume that initially all the edges are open and that, as suspension flows through the filter, edges clog, but different edges do not clog simultaneously.

We study our filter at a fixed time. In other words, when we say that an edge is clogged, we mean that the edge is clogged at that fixed time. Analogously, when we say that an edge is open, we mean open at that fixed time.

4.3. Mass conservation. In this subsection we introduce a definition and an observation that we will need later in the paper. This observation is a consequence of the law of mass conservation. Here, as in the rest of this section, G is a fixed multigraph that corresponds to one of our filters, such as the one in Figure 2.

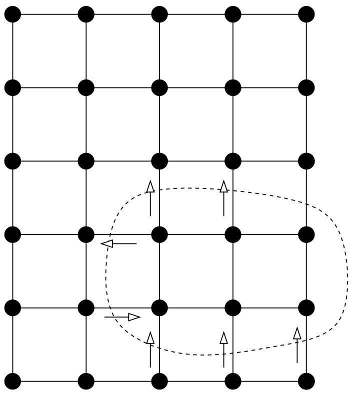


FIG. 3. *Ω is the bounded open set whose boundary, $\partial\Omega$, is in dashed lines. An arrow next to an edge that intersects $\partial\Omega$ indicates the direction of the flow within that edge. No arrow next to an edge that intersects $\partial\Omega$ indicates that there is not flow through that edge. E_Ω^{in} is the set of edges that have arrows next to them pointing into Ω. E_Ω^{out} is the set of edges that have arrows next to them pointing out of Ω.*

DEFINITION 4.2. *Let Ω be an open bounded set of $\mathbb{R}^2$ such that $\bar{\Omega}$, the closure of Ω, does not intersect any exterior edge of G, and $\partial\Omega$, the boundary of Ω, does not contain any node of G. We define*

$$E_\Omega = \{\textit{edges in } G \textit{ with exactly one end point in } \Omega\}, \tag{4.1}$$

$$E_\Omega^{\mathrm{in}} = \{e \in E_\Omega : \textit{suspension flows through } e \textit{ into } \Omega\}, \tag{4.2}$$

and

$$E_\Omega^{\mathrm{out}} = \{e \in E_\Omega : \textit{suspension flows through } e \textit{ out of } \Omega\}. \tag{4.3}$$

In Figure 3 we illustrate these definitions.

Since there may be some edges without flow through them, the union of E_Ω^{in} and E_Ω^{out} need not be E_Ω. In particular, clogged edges in E_Ω are neither in E_Ω^{in} nor in E_Ω^{out}. Note that there may also be open edges E_Ω without flow through them. Thus, there may be open edges in E_Ω that are in neither E_Ω^{in} nor E_Ω^{out}. Note also that E_Ω^{in} and E_Ω^{out} are disjoint sets.

OBSERVATION 4.1. *Let Ω be an open bounded set of $\mathbb{R}^2$ such that $\bar{\Omega}$ does not intersect any of the exterior edges of G and $\partial\Omega$ does not contain any node of G. Then we have the following:*

1. *The rate at which suspension flows into Ω through the edges in E_Ω^{in} is equal to the rate at which suspension flows out of Ω through the edges in E_Ω^{out}.*
2. *Let $e \in E_\Omega$. If all the other edges in E_Ω are clogged, then there is no flow through e.*
3. *If E_Ω is not empty, then at least one of the edges in E_Ω is not clogged.*

Proof. Suspension can flow into the filter only through its bottom boundary (i.e., the exterior nodes at $y = y_b$) and out of the filter only through its top boundary. Thus, since Ω does not contain any of the exterior nodes, there are neither mass sources nor mass sinks within Ω. This, together with the fact that the suspension is incompressible, implies point 1.

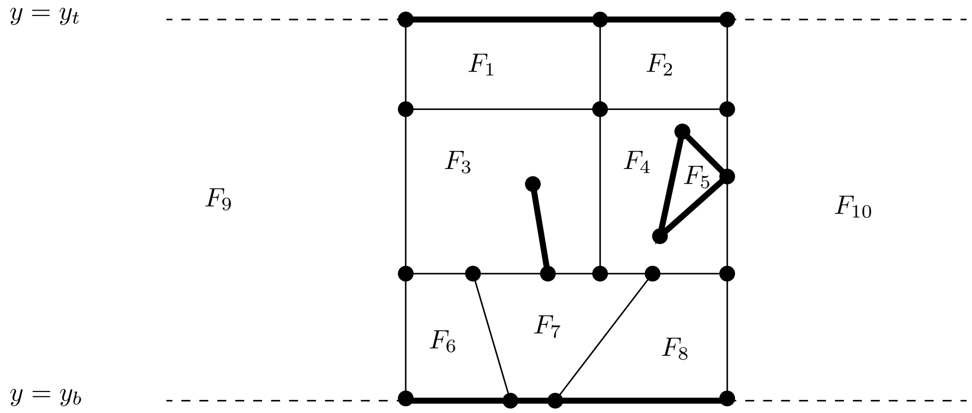


FIG. 4. *Example of a multigraph G. The edges of G are the thin and thick solid lines. The dashed lines are not part of G. F_i ($1 \leq i \leq 10$) are the connected components of $\{y_b < y < y_t\} - G$. The edges in thick solid lines are not contained in any percolating path. The edges in thin solid lines are contained in percolating paths. There can be flow only through the thin edges.*

Let e be an edge in E_Ω. If there is flow through e into Ω, point 1 implies that there should be flow out of Ω through some edge in E_Ω other than e. This is a contradiction since all the edges in E_Ω other than e are clogged. Thus, there is no flow through e into Ω. A similar argument shows that there is no flow through e out of Ω either, which proves point 2.

We prove point 3 by contradiction. Assume that all the edges in E_Ω are clogged. Let e be the edge in E_Ω that clogged last. Once the other edges were clogged, there was no more flow through e, and thus e could not have clogged, because our model assumes that an open edge does not clog if there is no flow through the edge. □

4.4. $C^\star$, a multigraph associated with the clogged edges. In this subsection we construct a multigraph $C^\star$ that is associated with the set of clogged edges.

We first note that the bounded connected components of the set $\{y_b < y < y_t\} - G$ are the bounded faces of G. In addition, $\{y_b < y < y_t\} - G$ has two unbounded connected components, one to the left of G and the other to its right. An example is shown in Figure 4, where the edges of G are the thin and thick solid lines. The dashed lines are not part of G. F_i ($1 \leq i \leq 10$) are the connected components of $\{y_b < y < y_t\} - G$. While F_i for $1 \leq i \leq 8$ are the bounded faces of G, F_9 and F_{10} are not faces of G.

Before proceeding with the construction of $C^\star$ we first need some preliminary definitions and observations.

DEFINITION 4.3. *We say that a path $P = n_0, e_1, n_1, \ldots, e_r, n_r$ in G is a percolating path if n_0 is a bottom exterior node, n_r is a top exterior node, and $n_1, \ldots, n_{r-1}$ are interior nodes.*

OBSERVATION 4.2. *Let e be an edge in G. If there is no percolating path that contains e, then e never clogs.*

The claim of Observation 4.2 is illustrated in Figure 4. Let e be an edge of G. If there is no percolating path that contains e, then the pressures at the end points of e are equal, which implies that there is no flow through e, and thus e can never clog.

Note that all percolating paths split the strip $\{y_b \leq y \leq y_t\}$ into two connected components, one to the right of the path and the other to its left.

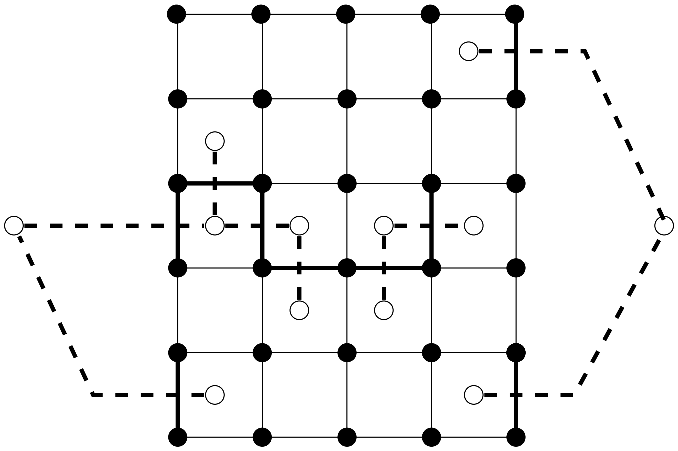


Fig. 5. *The clogged edges are thick solid lines. The open edges are in thin solid lines. The thick dashed lines are the edges in $C^\star$. The white circles are the nodes of $C^\star$.*

Definition 4.4. *Let P be a percolating path. We say that a set S is to the right of P if S is included in the closure of the right connected component of $\{y_b \leq y \leq y_t\} - P$. Analogously, S is to the left of P if S is included in the closure of the left connected component of $\{y_b \leq y \leq y_t\} - P$.*

For example, the set F_{10} in Figure 4 is to the right of any percolating path of the graph of that figure.

Observation 4.3. *Let P be a percolating path and F a connected component of $\{y_b < y < y_t\} - G$. Then F is either to the right of P or to the left of P.*

While obvious, the last observation leads to the next one that will be key in our construction of $C^\star$.

Observation 4.4. *Let e be an edge in G. If there is a percolating path that contains e, then e is in the boundary of two connected components of $\{y_b < y < y_t\} - G$.*

We are now ready to start our construction of a drawing of $C^\star$.

Select a point inside each connected component of the set $\{y_b < y < y_t\} - G$. We call this set of points $\mathcal{N}^\star$.

For each edge of G that is clogged, we draw exactly one edge of $C^\star$ as follows. Let e be a clogged edge of G. Observations 4.2 and 4.4 imply that e is included in the boundary of two connected components of $\{y_b < y < y_t\} - G$. Let $a^\star$ and $b^\star$ be the points of $\mathcal{N}^\star$ that are included in these components. We draw exactly one edge $e^\star$ of $C^\star$ connecting $a^\star$ and $b^\star$ in such a way that $e^\star$ intersects e in exactly one point, $e^\star$ does not intersect any other edge of G, and $e^\star \in \{y_b < y < y_t\}$. We say that $e^\star$ is the edge of $C^\star$ associated with e. This construction is carried out in such a way that edges of $C^\star$ may intersect only at their end points. The nodes of $C^\star$ are the end points of the edges in $C^\star$. Note that the set of nodes of $C^\star$ is a subset of $\mathcal{N}^\star$.

In Figure 5 we show an example of a set of clogged edges and the associated $C^\star$. The clogged edges are the thick solid lines, and the edges of $C^\star$ are the dashed lines. The white circles are the nodes of $C^\star$.

4.5. Bounding the number of clogged edges. The next sequence of observations will allow us to bound the number of clogged edges.

Observation 4.5. *The number of clogged edges is equal to the number of edges in $C^\star$.*

This last observation is an immediate consequence of the definition of $C^\star$.

Observation 4.6. *$C^\star$ does not have any bounded faces.*

Proof. The proof proceeds by contradiction. Assume that Ω is a bounded face of $C^\star$. From the definition of $C^\star$, the edges of G that intersect $C^\star$ are clogged. Thus, all the edges of G that intersect $\partial\Omega$ are clogged. Note also that $\bar{\Omega}$, the closure of Ω, does not intersect any of the exterior edges of G; $\partial\Omega$, the boundary of Ω, does not contain any node of G; and the number of edges of G that intersect $\partial\Omega$ is positive.

The above paragraph is in contradiction of point 3 of Observation 4.1. Thus, $C^\star$ does not have any bounded faces. □

As a consequence, the only face of $C^\star$ is its unbounded face. Thus, we have the following.

Observation 4.7. *$C^\star$ has only one face.*

Due to the definition of $C^\star$, we also have the following observation.

Observation 4.8. *Let $n_{C^\star}$ be the number of nodes of $C^\star$. Then, $n_{C^\star} \leq f_G + 1$, where f_G is the number of faces of G.*

We are now ready to bound the number of clogged edges. Let $n_{C^\star}$, $e_{C^\star}$, $f_{C^\star}$, and $\ell_{C^\star}$ be the number of nodes, edges, faces, and connected components of $C^\star$. Euler's formula implies

$$e_{C^\star} = n_{C^\star} + f_{C^\star} - \ell_{C^\star} - 1. \tag{4.4}$$

From Observation 4.7 we have $f_{C^\star} = 1$. Thus, (4.4) reduces to

$$e_{C^\star} = n_{C^\star} - \ell_{C^\star}. \tag{4.5}$$

As a consequence, using Observation 4.8, we have

$$e_{C^\star} \leq f_G + 1 - \ell_{C^\star}, \tag{4.6}$$

where f_G is the number of faces of G. Finally, since $\ell_{C^\star} \geq 1$ and $e_{C^\star}$ is the number of clogged edges, we obtain our bound, which we summarize in the following theorem.

Theorem 4.5. *Let G be a multigraph that corresponds to one of our filters. Then,*

$$\#\{\textit{clogged edges}\} \leq \#\{\textit{faces of } G\}. \tag{4.7}$$

5. Optimality of the bound.

As always, G is the multigraph of one of our filters.

Definition 5.1. *We say that $e_1, e_2, \ldots, e_s$ is a feasible clogging sequence or (for short) feasible sequence if, for each $1 \leq i \leq s$, there is flow through the edge e_i when $e_1, e_2, \ldots, e_{i-1}$ are clogged and all the other edges are open. We say that s is the length of the sequence.*

Recall that an edge can clog only when suspension flows through it. Thus, if q edges clogged, and the ith edge that clogged was e_i, then $e_1, e_2, \ldots, e_q$ is a feasible sequence of edges. Note that the bound of section 4 is actually a bound on the length of feasible sequences of edges.

While there are many feasible sequences that make the filter nonpermeable, only one actually realizes. The flow conditions, conductivity of the channels, as well as other factors determine the feasible sequence that realizes, which generally has fewer edges than other feasible clogging sequences. It is not our goal to find the sequence that realizes. In this section we show that, if every interior edge of G is contained in a

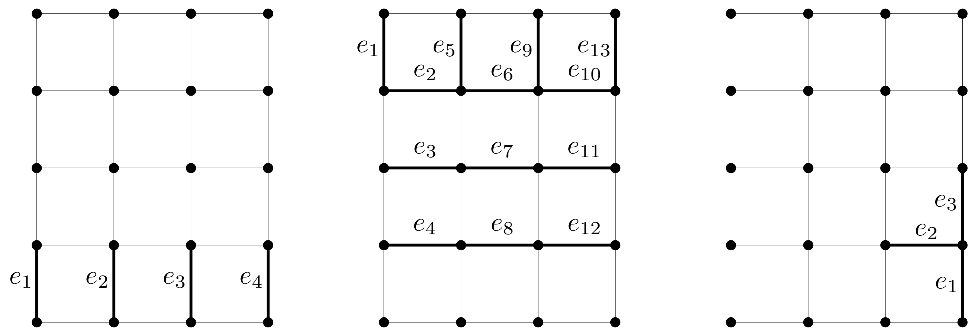


Fig. 6. *The sequences $e_1, \ldots, e_r$ in the left and middle figures ($r = 4$ in the left figure and $r = 13$ in the middle figure) are two feasible sequences of edges that make the filter nonpermeable. Note that the number of faces of the graph is 13. Thus, the sequence of the middle figure is of maximum length. Our work does not predict whether the sequence in the left figure, the one in the middle, or another sequence is realized; thus, we can predict only that the sequence that realizes has length less than or equal to 13. The sequence in the right figure is not feasible. Once e_1 and e_2 clog, there is no more flow through e_3, and thus it cannot clog.*

percolating path, our bound is sharp in the sense that there exists a feasible sequence of edges whose length is equal to our bound, the number of faces of G (the right-hand side of (4.7)). However, it should be noted that the length of the feasible sequence that actually realizes and makes the filter nonpermeable may be smaller. In other words, while the length of the longest feasible sequence of edges is an upper bound on the number of channels that actually clog, these numbers may not be equal. Illustrative examples are given in Figure 6. We will come back to this issue in section 7.

Let e be an interior edge of G. As illustrated in Figure 4 and previously discussed, if there is no percolating path that contains e, then the pressure at the end points of e are equal, which implies that there is no flow through e and thus that e can never clog. Removing first all such edges from G, then all the exterior edges, then the nodes that are left isolated, and finally adding new exterior edges as necessary leads to a new multigraph $\tilde{G}$ for which the bound will be attained. Note that the flow in G is exactly equal to the flow in $\tilde{G}$. There is no flow within edges of G that do not belong to $\tilde{G}$.

5.1. Leftmost percolating paths.

Observation 5.1. *Let $P = n_0, e_1, n_1, \ldots, e_r, n_r$ be a percolating path. Let e be an edge to the left of P. If e is included in a percolating path, then there exists a percolating path R such that P and e are to the right of R.*

Proof. Assume that e is not in P, since otherwise the observation is trivially true by selecting $R = P$. Let Q be a percolating path that contains e. Let $\bar{Q}$ be the largest path that satisfies (1) $\bar{Q}$ is included in Q, (2) $\bar{Q}$ contains e, (3) $\bar{Q}$ is to the left of P, and (4) $\bar{Q}$ may intersect P only at the end points of $\bar{Q}$.

If $\bar{Q} = Q$, as in Figure 7(a), select $R = Q$. Note that e and P are to the right of R. Otherwise, $\bar{Q}$ intersects P. In this case, $\bar{Q} \cap P$ splits P into two or three connected sections. Replacing one of these connected sections with $\bar{Q}$ leads to the percolating path R we are looking for. If $\bar{Q}$ contains a bottom exterior node, as in Figure 7(b), we replace the section of P that has a bottom exterior node. If $\bar{Q}$ contains a top exterior node, as in Figure 7(c), we replace the section of P that has a top exterior node. If $\bar{Q}$ does not contain any exterior node, as in Figure 7(d), we replace the section of P

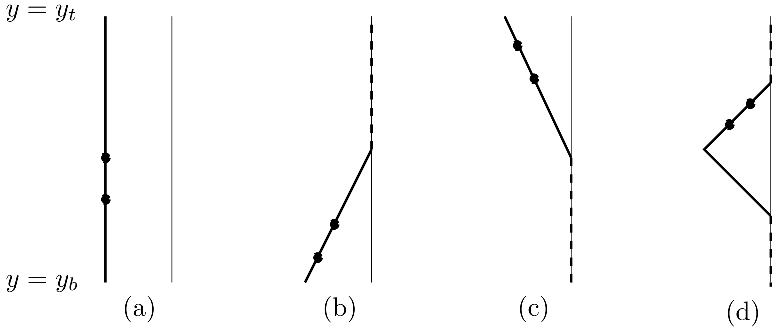


FIG. 7. *The four different possibilities of Observation* 5.1. *We do not show all the multigraph* G*, only* e*,* $\bar{Q}$*, and* P*. The edge* e *is the segment between the solid small circles.* P *is the thin solid vertical line,* $\bar{Q}$ *the thick solid line, and* R *the union of* $\bar{Q}$ *and the dashed lines.*

without exterior nodes. □

This last observation and the fact that each exterior node is the end point of an interior edge lead to the following.

OBSERVATION 5.2. *If every interior edge in* G *is included in a percolating path, then there is a unique percolating path* P *in* G *such that* G *is to the right of* P*. We call* P *the leftmost percolating path of* G*.*

5.2. A feasible sequence of maximum length. The first edge. Assume that every interior edge in G is included in a percolating path. Our goal is to construct a feasible sequence of edges $e_1, e_2, \ldots, e_N$ of maximum length. Let P be the leftmost percolating path of G. In this subsection we identify $\bar{P}$, a subpath of P, from which e_1 will be selected. The selection of $\bar{P}$ is done with care so that the rest of the sequence, $e_2, \ldots, e_N$, can be constructed inductively, as we will do in the next subsection.

OBSERVATION 5.3. *Assume that every interior edge in* G *is included in a percolating path of* G*. Let* P *be the leftmost percolating path of* G*. Assume that there are no bounded faces* F *of* G *such that* $\partial F \cap P$ *contains an edge. Then* $P = G$*.*

Proof. The proof is by contradiction. Assume $P \neq G$. Then there exists an edge in G not in P. In fact, since every exterior node is the end point of an interior edge, we have that there exists an interior edge e in G such that e is not in P. Let Q be a percolating path in G containing e. Note that there is a nonempty open bounded Ω enclosed by P, Q, $y = y_b$, and $y = y_t$. Note also that at least one edge of P is in the boundary of Ω. The closure of Ω is the union of the closure of the bounded faces of G included in Ω. Thus, there exists F, a bounded face of G, such that $\partial F \cap P$ contains an edge. This is a contradiction, which proves the observation. □

OBSERVATION 5.4. *Assume that every interior edge in* G *is included in a percolating path of* G*. Let* F *be a bounded face of* G*. Then,* $\partial F \cap \{y = y_b\}$ *is connected, and* $\partial F \cap \{y = y_t\}$ *is also connected.*

Proof. Assume that $\partial F \cap \{y = y_b\}$ is not connected. Then, as illustrated in Figure 8, there is a path $\bar{Q}$ that is included in ∂F such that $\bar{Q} \cap \{y = y_b\}$ are the end points of $\bar{Q}$, the edges in $\bar{Q}$ are interior edges, and none of them is included in a percolating path of G, which contradicts our assumption. Thus, $\partial F \cap \{y = y_b\}$ is connected. Analogously, $\partial F \cap \{y = y_t\}$ is also connected. □

OBSERVATION 5.5. *Assume that every interior edge in* G *is included in a percolating path of* G*. Let* P *be the leftmost percolating path of* G*. Assume that* G *has a*

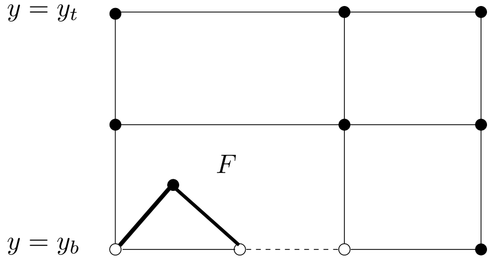


FIG. 8. *Multigraph G. $\partial F \cap \{y = y_b\}$ is the edge in dashed line and the white nodes, which is a disconnected set. The thick solid line is $\bar{Q}$. The edges in $\bar{Q}$ are not included in any percolating path of G.*

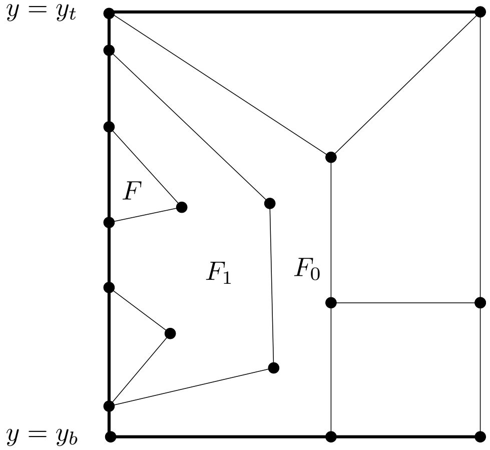


FIG. 9. *Multigraph G. The edges of the path S are in thick lines. Example of a sequence $F_0, F_1, F_2 = F$ constructed as in Observation* 5.5.

bounded face. Then, there exists F, a bounded face of G, such that

1. *$\partial F \cap P$ contains an edge, and*
2. *$\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ is connected.*

Proof. Let S be the path that results from the following steps. We start at the rightmost exterior node of the bottom boundary and walk left along that boundary toward the path P. We continue walking through P to the top boundary. We then walk right along the top boundary and end the path at the rightmost exterior node of the top boundary (see Figure 9).

Since G has a bounded face, $P \neq G$, and thus, from Observation 5.3, there exists F_0, a bounded face of G, such that $\partial F_0 \cap P$ contains an edge. Let $S_{F_0} = \partial F_0 \cap S$ and $\tilde{S}_{F_0}$ be the smallest path included in S that contains S_{F_0}. Recall that paths are connected. Thus, if $\tilde{S}_{F_0} = S_{F_0}$, then $F = F_0$ is a face we are looking for.

We next show that $P \cap \tilde{S}_{F_0} \subseteq S_{F_0}$ implies that $\tilde{S}_{F_0} = S_{F_0}$. Assume that $P \cap \tilde{S}_{F_0} \subseteq S_{F_0}$. If both $\partial F_0 \cap \{y = y_b\}$ and $\partial F_0 \cap \{y = y_t\}$ are empty sets, then $\tilde{S}_{F_0} = P \cap \tilde{S}_{F_0} \subseteq S_{F_0}$. Assume now that $\partial F_0 \cap \{y = y_b\}$ is nonempty. Thus, $\tilde{S}_{F_0}$ contains the exterior node of P in the bottom boundary, and since $P \cap \tilde{S}_{F_0} \subseteq S_{F_0}$, that exterior node is also in S_{F_0}. Thus, $\{y = y_b\} \cap S_{F_0} = \partial F_0 \cap \{y = y_b\}$ also contains that exterior node and, given that $\partial F_0 \cap \{y = y_b\}$ is connected due to Observation 5.4,

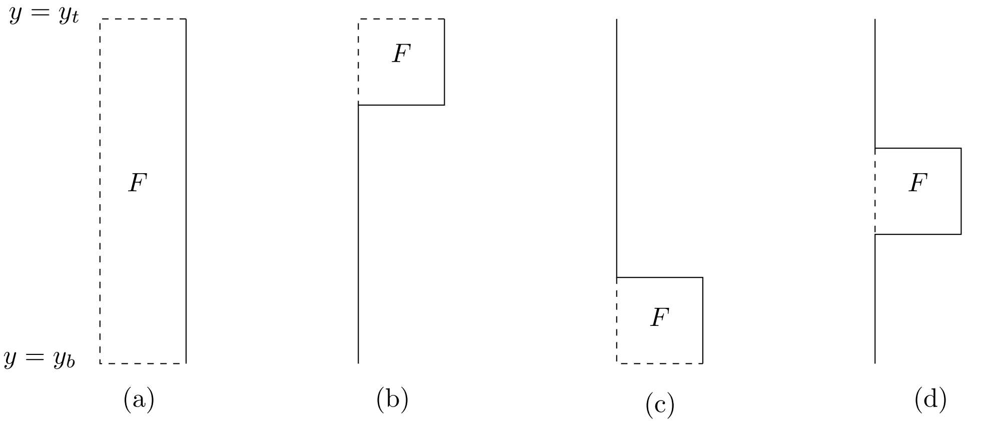


FIG. 10. *Four different possibilities of F from Observation 5.6 and Theorem 5.2. The solid line is the percolating path S of Observation 5.6. The left vertical line is P. $P \cap S$ is in solid line, and the section of P that does not intersect S is in dashed line. The dashed horizontal lines are the sections of the top and bottom boundary that are to the left of S.*

we have that $\{y = y_b\} \cap \tilde{S}_{F_0} \subseteq S_{F_0}$. This argument applied to the top boundary leads to $\tilde{S}_{F_0} = (P \cup \{y = y_t\} \cup \{y = y_b\}) \cap \tilde{S}_{F_0} \subseteq S_{F_0}$ if $(P \cap \tilde{S}_{F_0}) \subseteq S_{F_0}$.

We are left to show that the observation is true when $\tilde{S}_{F_0} \neq S_{F_0}$ (see Figure 9), and so we now assume $\tilde{S}_{F_0} \neq S_{F_0}$. Given the above paragraph, we have that $\tilde{S}_{F_0} - S_{F_0}$ intersects P in at least one edge, say e. As illustrated in Figure 10, we can select F_1, a bounded face of G, such that ∂F_1 contains e. Let $S_{F_1} = \partial F_1 \cap S$ and $\tilde{S}_{F_1}$ be the smallest path included in S that contains S_{F_1}. If $\tilde{S}_{F_1} = S_{F_1}$, then $F = F_1$ is a face we are looking for. Otherwise, we note that $\tilde{S}_{F_1}$ is included in a connected component of $\tilde{S}_{F_0} - S_{F_0}$, and thus $\tilde{S}_{F_1} \subsetneq \tilde{S}_{F_0}$. As a consequence, since G is a finite multigraph, repeating this procedure as many times as necessary, we will find the face F that we are looking for (see Figure 9). □

Assume that every interior edge in G is included in a percolating path. Let F be a bounded face of G that satisfies the conditions of Observation 5.5. Let $\tilde{P} = P \cap \partial F$, where P is the leftmost percolating path of G. In the next subsection we will show how to construct a feasible sequence of edges $e_1, e_2, \ldots, e_N$ of maximum length where e_1 will be selected from $\tilde{P}$.

5.3. Feasible sequence of edges of maximal length.

OBSERVATION 5.6. *Assume that every interior edge in G is included in a percolating path. Let P be the leftmost percolating path of G. Assume that G has a bounded face. Let F be a face of G such that (1) $\partial F \cap P$ contains an edge, and (2) $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ is connected.*

Then, for every interior edge e in G not in $\partial F \cap P$ there exists a percolating path Q of G such that Q contains e and Q does not have any edge in $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$.

Proof. Let S_1 be the section of P that does not intersect ∂F, i.e., $S_1 = P - \partial F$. Let S_2 be the section of ∂F that intersects neither P nor the boundaries, i.e., $S_2 = \partial F - (P \cup \{y = y_t\} \cup \{y = y_b\})$. Let S be the percolating path that results from walking along $S_1 \cup S_2$ (see Figure 10). Note that every edge in G not in $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ is to the right of S.

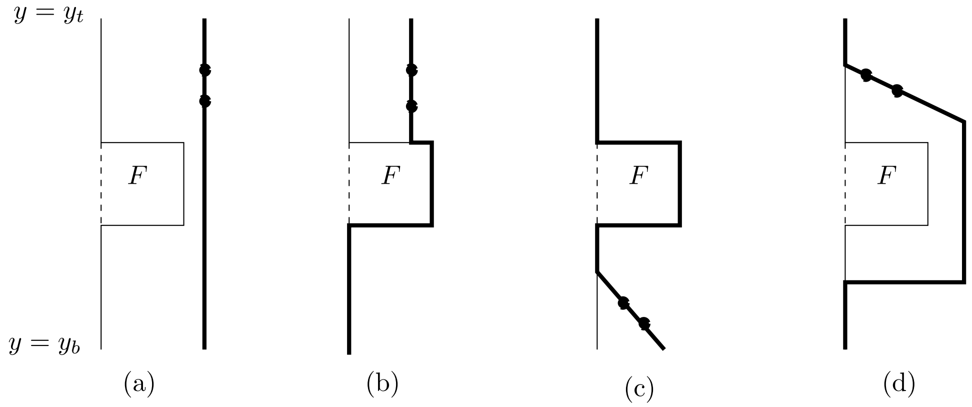


Fig. 11. *Four different possibilities of the percolating path Q, in thick solid lines, from Observation* 5.6. *The edge e is between the solid small circles. The left vertical line is P.*

Let e be an interior edge in G not in $\partial F \cap P$. Let R be a percolating path of G that contains e. Let $\bar{Q}$ be the largest subpath of R that contains e such that $\bar{Q}$ may only intersect S at the end points of $\bar{Q}$. As illustrated in Figure 11, we can construct a percolating path Q that contains $\bar{Q}$, may contain sections of S, but does not contain any edge outside $\bar{Q} \cup S$. Thus, Q contains e and does not have any edge in $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$. □

Observation 5.7. *Assume that every interior edge in G is included in a percolating path of G. Let P be the leftmost percolating path of G. Assume that G has a bounded face. Let F be a bounded face of G such that* (1) *$\partial F \cap P$ contains an edge, and* (2) *$\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ is connected. Let G' be the multigraph that results from first removing from G the edges in $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ and then removing the nodes that are left isolated. Then the following hold:*

1. *Every interior edge in G' is included in a percolating path in G'.*
2. *$f_{G'} = f_G - 1$; i.e., the number of faces of G' is equal to the number of faces of G minus one.*
3. *G' is the multigraph of one of our filters.*

Proof. Point 1 is an immediate consequence of Observation 5.6.

Point 2 results from the simple facts that (1) all the bounded faces of G' are bounded faces of G, (2) the only bounded face of G that is not a face of G' is F, and (3) both G and G' (as well as any multigraph) have only one unbounded face. (Note that G' results from removing the dashed lines in Figure 10.)

Point 3 is also clear. □

Theorem 5.2. *If every interior edge in G is included in a percolating path of G, then there exists a feasible sequence of edges of length f_G. Thus, our bound is optimal for this class of filters.*

Proof. We will prove the theorem by induction on f_G. First note that $f_G = 1$ if and only if G is a percolating path. In this case, any edge of G forms a feasible sequence of edges of length $f_G = 1$.

Assume now that $f_G > 1$. Let P be the leftmost percolating path of G. Note that P cannot be equal to G, since otherwise f_G would be equal to one.

Let F be a bounded face of G such that (1) $\partial F \cap P$ contains an edge, and (2) $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ is connected. Such a face exists by the observations of this section. Let e_1 be any edge in $\partial F \cap P$.

Let G' be the multigraph that results from first removing from G the edges in $\partial F \cap (P \cup \{y = y_t\} \cup \{y = y_b\})$ and then removing the nodes that are left isolated. From Observation 5.7, every interior edge in G' is included in a percolating path in G', $f_{G'} = f_G - 1$, and G' is the multigraph of one of our filters.

By inductive hypothesis there exists in G' a feasible sequence of edges of length $f_{G'} = f_G - 1$. For convenience, call one such sequence $e_2, \ldots, e_{f_G}$. From the observations of this section, it follows that $e_1, e_2, \ldots, e_{f_G}$ is a feasible sequence of edges in G, which proves the theorem. □

6. The bound in terms of the average degree of G for large filters. As always, G is a multigraph of one of our filters. We assume in this section that every interior edge of G is contained in a percolating path. We recall that f_G, e_G, and n_G are the numbers of faces, edges, and nodes of G, respectively. We also recall that d_G, the average degree of G, is given by $d_G = 2e_G/n_G$ (see (3.2)).

Assume that the number of edges is large, i.e., $e_G \gg 1$. In this case, the Euler formula $f_G + n_G = e_G + \ell_G + 1$ reduces to $f_G + n_G \approx e_G$ since $\ell_G = 1$. Thus, from (3.2), we have $f_G + 2e_G/d_G \approx e_G$. This leads to the following observation.

OBSERVATION 6.1. *If $e_G \gg 1$ and $d_G \neq 2$, then our bound* (4.7) *reads*

$$\#\{\textit{clogged edges}\} \lesssim \frac{d_G - 2}{d_G} e_G. \tag{6.1}$$

In particular, if every interior edge of G is contained in a percolating path, the number of edges in a feasible sequence of edges with maximum length is asymptotically $((d_G - 2)/d_G)e_G$.

In many situations of interest, G is a graph; i.e., no two edges have the same end points. For example, if all the edges in a multigraph are straight segments, then the multigraph is really a graph. It is a well-known fact from graph theory that, if G is a planar graph, the average degree of G is bounded by 6, i.e., $d_G \leq 6$. This leads to the following observation.

OBSERVATION 6.2. *If G is a graph and $e_G \gg 1$, then*

$$\#\{\textit{clogged edges}\} \lesssim \frac{2}{3} \#\{\textit{all edges}\}. \tag{6.2}$$

A natural goal is to design filters that use as much of the pore space as possible to trap particles before the filter ceases to be permeable. Thus, of particular interest is to know the proportion of channels that are clogged when the filter ceases to be permeable. The last observation provides a bound on this quantity whenever G is a graph.

As particular examples, consider the graphs of Figure 12. At this point do not make a distinction between solid and dashed edges. In the large filter limit, i.e., the distance between the top and bottom boundaries is much larger than the length of the edges, the graph with square bounded faces satisfies $d_G \approx 4$, and thus, for this graph, (6.1) implies $\#\{\text{clogged edges}\} \lesssim e_G/2$. For the graph with triangle bounded faces, we have $d_G \approx 6$ and thus $\#\{\text{clogged edges}\} \lesssim 2e_G/3$. For the graph with hexagonal bounded faces, we have $d_G \approx 3$ and thus $\#\{\text{clogged edges}\} \lesssim e_G/3$.

7. A subclass of filters and examples. In this section, we consider filters in which every interior edge is included in a percolating path. We have shown that, for this kind of filters, our bound is sharp; i.e., there is a feasible sequence of edges whose number of edges or length is equal to our bound, i.e., the number of faces of

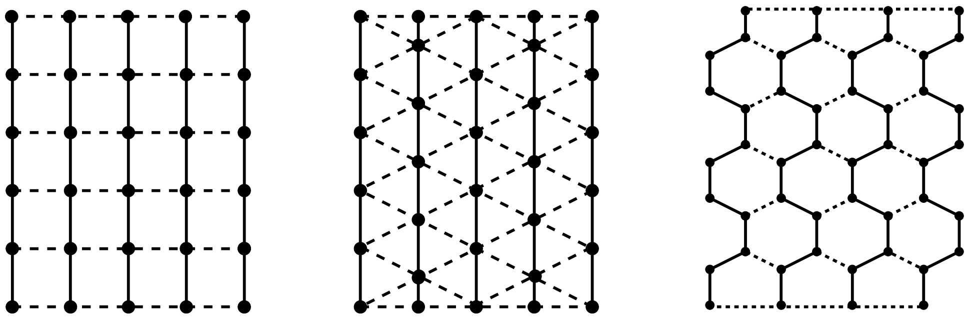


FIG. 12. *Multigraphs with geometries of class $\mathcal{A}$. The percolating paths P_i are in solid lines. The crossing paths $H_{i,j}$ are in dashed lines.*

the multigraph (see the right-hand side of (4.7)). However, in general, there are many feasible sequences of edges that make the filter nonpermeable, and the length of most of them is less than our bound. Thus, the feasible sequence of edges that is realized, i.e., the sequence of edges that actually clog and make the filter nonpermeable, which depends on the flow conditions as well as the width of the channels, is, in general, much shorter than the feasible sequences of edges with maximum length.

In this section, we will restrict our attention to a subclass of filters for which we will show how to select the width of the channels so that, as the suspension flows, the feasible sequence of edges that is realized and makes the filter nonpermeable will indeed have maximum length.

7.1. Subclass of filters. The geometries.

DEFINITION 7.1. *We say that a multigraph G that corresponds to one of our filters has geometry of class $\mathcal{A}$ if there is a nonnegative integer r and a sequence of disjoint percolating paths $P_0, P_1, \ldots, P_r$ in G such that, for each i, $0 \leq i < r$, the following hold:*

1. *P_i is to the left of P_{i+1}.*
2. *There are positive integers s_i such that P_i and P_{i+1} are joined by $s_i + 1$ paths that may have only end points in common; i.e., for each j, $0 \leq j \leq s_i$, there is a path $H_{i,j}$ such that (1) $H_{i,j} \cap P_i$ is an end point of $H_{i,j}$, (2) $H_{i,j} \cap P_{i+1}$ is the other end point of $H_{i,j}$, and (3) H_{i,j_1} and H_{i,j_2} may intersect at their end points only if $j_1 \neq j_2$.*
3. *G is the union of the percolating paths P_i $(0 \leq i \leq r)$ and the "crossing" paths $H_{i,j}$ $(0 \leq i < r,\ 0 \leq j \leq s_i)$.*

For convenience, we assume that the paths $H_{i,j}$ are labeled in such a way that $H_{i,j+1}$ is above $H_{i,j}$ $(0 \leq i < r,\ 0 \leq j < s_i)$. More precisely, $H_{i,j+1}$ is in the closure of the bounded region whose boundary is included in $\{y = y_t\} \cup P_i \cup P_{i+1} \cup H_{i,j}$. Note that $H_{i,0}$ is included in the bottom boundary and H_{i,s_i} is included in the top boundary.

Examples of multigraphs that have the geometry of class $\mathcal{A}$ are shown in Figure 12. The percolating paths P_i are in solid lines. The crossing paths $H_{i,j}$ are in dashed lines.

7.2. Subclass of filters. The width of the channels. The physical mechanisms that lead to the clogging of channels may be complex and depend on the particular problem under consideration. Here we will assume the following simple rules. Each channel is either thin or thick. Thick channels never clog, and thin channels eventually clog if there is flow through them.

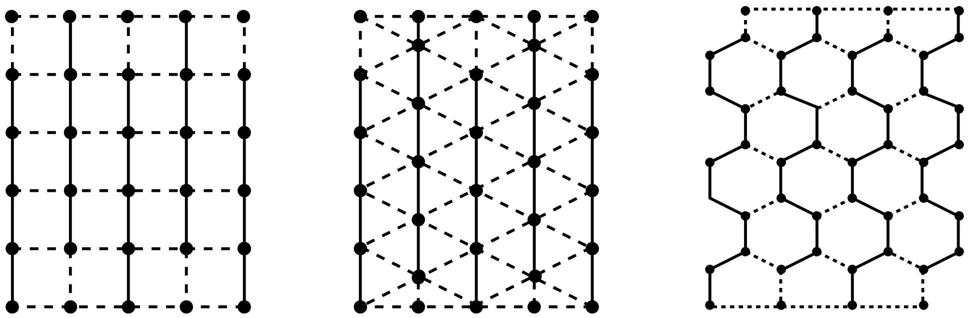


FIG. 13. *Multigraphs of Figure* 12. *The thick edges are in solid lines. Thin edges in dashed lines.*

We now select the thin and thick edges. Let G be a multigraph that has geometry of class $\mathcal{A}$. Let $P_0, P_1, \ldots, P_r$ be the percolating paths as in Definition 7.1. Every percolating path is split into subpaths by the crossing paths. For i even, let $\bar{P}_i$ be the subpath of P_i that contains an exterior node at the top boundary. For i odd, let $\bar{P}_i$ be the subpath of P_i that contains an exterior node at the bottom boundary. One edge in each of the subpaths $\bar{P}_i$ for any i is selected to be thin. One edge in each of the crossing paths $H_{i,j}$ is also selected to be thin. Every other edge is chosen thick. In Figure 13 we show the multigraphs of Figure 12, but now the thick edges are in solid lines and the thin edges in dashed lines.

DEFINITION 7.2. *We say that a multigraph G that corresponds to one of our filters is of class $\mathcal{A}$ if G has geometry of class $\mathcal{A}$ and the width of the edges of G are selected as described above.*

7.3. The bound realizes for filters of class $\mathcal{A}$. We now show that, for the filters considered in this section, i.e., with multigraph of class $\mathcal{A}$, the bound realizes; i.e., the number of edges that actually clog is equal to our bound, the number of faces of the multigraph of the filter. We show this in two steps. We first show that in each of the paths $\bar{P}_i$ the thin edge clogs, and in each of the paths $H_{i,j}$ not included in the top or bottom boundaries, i.e., $H_{i,j}$ for $1 \leq j < s_i$, the thin edge also clogs. Then, we show that the number of these paths is equal to the number of faces of the multigraph of the filter.

OBSERVATION 7.1. *In each of the paths $\bar{P}_i$ ($0 \leq i \leq r$) the thin edge clogs, and in each of the paths $H_{i,j}$ not included in the top or bottom boundaries ($0 \leq i < r$, $1 \leq j < s_i$) the thin edge clogs.*

Proof. We first note that there is no percolating paths with all thick edges. Thus, the filter eventually ceases to be permeable.

Let e_i be the thin edge in $\bar{P}_i$. All the other edges in P_i are thick, and thus they never clog. As a consequence, while e_i is open, the filter is permeable. Thus, e_i eventually clogs.

Let $H_{i,j}$ be one of the crossing paths not included in the top or bottom boundaries, i.e., $1 \leq j < s_i$. We connect one end point of $H_{i,j}$ to a bottom exterior node and the other to a top exterior node with paths of thick edges as follows. Let Q be the subpath of P_i that has an exterior node as an end point, shares the other end point with $H_{i,j}$, and all the edges in Q are thick. Let R be the subpath of P_{i+1} that has an exterior node as an end point, shares the other end point with $H_{i,j}$, and all the edges in R are thick. From the discussion of sections 7.1 and 7.2, it is clear that Q and R are well defined. It is also clear that the union of Q, R, and $H_{i,j}$ forms a percolating

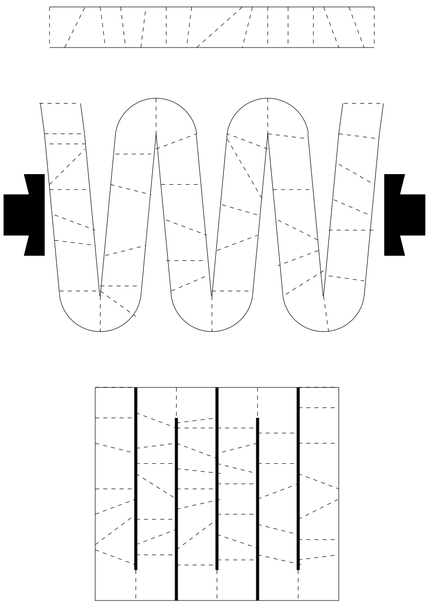


FIG. 14. *Building a filter of class $\mathcal{A}$. In the top figure we start with a thin filter with thin channels. In the middle figure we fold the thin filter. The bottom figure shows the resulting filter after folding and compressing. The thick edges are in solid thick lines. The thin edges are in dashed lines.*

path whose only thin edge is the one in $H_{i,j}$. Due to the same argument of the last paragraph, this implies that the thin edge in $H_{i,j}$ eventually clogs. ∎

OBSERVATION 7.2. *The number of the paths $\bar{P}_i$ ($0 \leq i \leq r$) and $H_{i,j}$ not included in the top or bottom boundaries ($0 \leq i < r$, $1 \leq j < s_i$) is equal to f_G, the number of faces of G.*

Proof. We first note that the number of the paths $\bar{P}_i$ ($0 \leq i \leq r$) and $H_{i,j}$ not included in the top or bottom boundaries ($0 \leq i < r$, $1 \leq j < s_i$) is equal to $1 + r + \sum_{i=0}^{r-1}(s_i - 1) = 1 + \sum_{i=0}^{r-1} s_i$. Thus our goal reduces to showing that $f_G = 1 + \sum_{i=0}^{r-1} s_i$.

Let $0 \leq i < r$ and $1 \leq j \leq s_i$. If we remove from the plane the paths P_i, $H_{i,j-1}$, P_{i+1}, and $H_{i,j}$, we are left with one bounded and one unbounded connected component. Let $F_{i,j}$ be the bounded component. It is clear that the bounded faces of G are $F_{i,j}$ for $0 \leq i < r$ and $1 \leq j \leq s_i$. Therefore, $f_G = 1 + \sum_{i=0}^{r-1} s_i$, which completes the proof. ∎

7.4. Building filters of class $\mathcal{A}$ from thin filters. We now discuss a possible means, at least theoretically, to construct filters of class $\mathcal{A}$.

We start with a thin filter with thin channels. By a thin filter we mean that the corresponding graph is the union of disjoint percolating paths connecting the top and bottom boundaries. An example is shown in the top figure of Figure 14.

Next, we fold the thin filter as shown in the middle figure of Figure 14. As we compress the folded filter from the sides, the spaces between folds become the thick channels, and we are left with the new filter shown in the bottom figure of Figure 14. The applied pressure in the folding step should be high, but not so high so the new pore space, the thick channels, are in fact thicker than the channels in the original thin filter before folding. The resulting filter is a filter of class $\mathcal{A}$.

Note that the above discussion suggests a way to construct filters of *long life* even if the original thin filter does not strictly satisfy the condition of having the corresponding graph be the union of disjoint percolating paths.

7.5. Further comments on filters of class $\mathcal{A}$. As channels clog, the permeability of the filter decreases. This is unavoidable. Nevertheless, we expect that this decrease in permeability will be relatively slow (as compared with other filters) for filters of class $\mathcal{A}$, because the suspension can flow with relative ease along the thick channels and, as shown in the proof of Observation 7.1, this family of filters has lots of percolating paths where all but one edge are thick.

REFERENCES

[1] S. L. Bryant, D. W. Mellor, and C. A. Cade, *Physically representative network models of transport in porous media*, AIChE J., 39 (1993), pp. 387–396.

[2] V. N. Burganos, C. A. Paraskeva, and A. C. Payatakes, *Three-dimensional trajectory analysis and network simulation of deep bed filtration*, J. Colloid Interf. Sci., 148 (1992), pp. 167–181.

[3] H. C. Chan, S. C. Chen, and Y. I. Chang, *Simulation: The deposition behavior of Brownian particles in porous media by using the triangular network model*, Sep. Purif. Technol., 44 (2005), pp. 103–114.

[4] F. Civan, *Reservoir Formation Damage*, Gulf Publishing Company, Houston, TX, 2000.

[5] G. Dagan, *Flow and Transport in Porous Formations*, Springer-Verlag, Berlin, 1989.

[6] S. Datta and S. Redner, *Gradient clogging in depth filtration*, Phys. Rev. E, 58 (1998), pp. 1203–1206.

[7] R. Diestel, *Graph Theory*, 3rd ed., Springer, Berlin, 2005.

[8] E. C. Donaldson, B. A. Baker, and H. B. Carrol, *Particle transport in sandstones*, SPE paper 6905, presented at the 52nd annual fall meeting of the SPE of AIME (Denver, CO), 1977.

[9] F. A. L. Dullien, *Porous Media. Fluid Transport and Pore Structure*, 2nd ed., Academic Press, New York, 1992.

[10] I. Fatt, *The network model of porous media*—I. *Capillary pressure characteristics*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 144–159.

[11] I. Fatt, *The network model of porous media*—II. *Dynamic properties of a single size tube network*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 160–163.

[12] I. Fatt, *The network model of porous media*—III. *Dynamic properties of networks with tube radius distributions*, Trans. Am. Inst. Min. Engrs., 207 (1956), pp. 164–181.

[13] G. Grimmett, *Percolation*, 2nd ed., Springer, Berlin, 1999.

[14] J. H. D. Hampton, S. B. Savage, and R. A. L. Drew, *Computer modeling of filter pressing and clogging in a random tube network*, Chem. Eng. Sci., 48 (1993), pp. 1601–1611.

[15] J. P. Herzig, D. M. Leclerc, and P. Le Goff, *Flow of suspensions through porous media—Application to deep filtration*, Ind. Eng. Chem., 62 (1970), pp. 8–35.

[16] A. O. Imdakm and M. Sahimi, *Transport of large particles in flow through porous media*, Phys. Rev. A, 36 (1987), pp. 5304–5309.

[17] A. O. Imdakm and M. Sahimi, *Computer-simulation of particle-transport processes in flow through porous media*, Chem. Eng. Sci., 46 (1991), pp. 1977–1993.

[18] J. M. Montgomery, *Water Treatment Principles and Design*, John Wiley & Sons, New York, 1985.

[19] K. C. Khilar and H. S. Fogler, *Migration of Fines in Porous Media*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[20] Y. S. Kim and A. J. Whittle, *Filtration in a porous granular medium:* 2. *Application of bubble model to* 1-*D column experiments*, Transp. Porous Media, 65 (2006), pp. 309–335.

[21] J. Lee and J. Koplik, *Network model for deep bed filtration*, Phys. Fluids, 13 (2001), pp. 1076–1086.

[22] B. E. Logan, *Environmental Transport Processes*, John Wiley & Sons, New York, 1999.

[23] J. D. Logan, *Transport Modeling in Hydrogeochemical Systems*, Springer-Verlag, New York, 2001.

[24] L. M. McDowell-Boyer, J. R. Hunt, and N. Sitar, *Particle transport through porous media*, Water Resource Res., 22 (1986), pp. 1901–1921.

[25] S. D. Rege and H. S. Fogler, *A network model for deep bed filtration of solid particles and emulsion drops*, AIChE J., 34 (1988), pp. 1761–1772.

[26] J. N. Ryan and M. Elimelech, *Colloid mobilization and transport in groundwater*, Colloids Surfaces A, 107 (1996), pp. 1–56.

[27] M. Sahimi, *Applications of Percolation Theory*, Taylor & Francis, London, 1994.

[28] M. Sahimi, G. R. Gavalas, and T. T. Tsotsis, *Statistical and continuum models of fluid-solid reactions in porous media*, Chem. Eng. Sci., 45 (1990), pp. 1443–1502.

[29] B. J. Suchomel, B. M. Chen, and M. B. Allen, *Network model of flow, transport and biofilm effects in porous media*, Transp. Porous Media, 30 (1998), pp. 1–23.

[30] K. E. Thompson and H. S. Fogler, *Modeling flow in disordered packed beds from pore-scale fluid mechanics*, AIChE J., 43 (1997), pp. 1377–1389.

[31] D. Tiab and E. C. Donaldson, *Petrophysics*, Gulf Publishing Company, Houston, TX, 1996.

[32] C. Tien and A. Payatakes, *Advances in deep bed filtration*, AIChE J., 25 (1979), pp. 737–759.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 763–768

SHORT WAVE STABILITY FOR INVISCID SHEAR FLOW*

MICHAEL RENARDY†

Abstract. We consider the linear stability of inviscid shear flows. While it is well known that discontinuous velocity profiles lead to short wave instabilities and ill-posedness, known examples of instability for smooth profiles have a short wave cutoff; i.e., there is a critical wave number beyond which no unstable eigenvalues exist. This paper proves a result to this effect under suitable assumptions on the base flow profile.

Key words. hydrodynamic stability, Rayleigh equation, inviscid shear flow

AMS subject classification. 76E05

DOI. 10.1137/080720905

1. Introduction.

The linear stability of inviscid shear flows has been studied extensively over the past century. I refer to [1, 2, 3] for reviews. Discontinuous velocity profiles lead to the Kelvin–Helmholtz instability which has unbounded growth rates in the limit of high wave numbers. In contrast, the instabilities in shear flows with smooth velocity profiles appear to be long wave instabilities. That is, there is cutoff at some maximal wave number α_m such that there are no unstable eigenvalues for $\alpha > \alpha_m$.

Although all known examples appear to satisfy this, I have not been able to find a proof in the literature. Almost half a century ago, Howard [4] proved such a result for a special case. He assumes that all inflection points of the base flow profile U occur at the same value U_i and that $U''/(U - U_i)$ is bounded and of one sign.

The objective of this paper is to prove the nonexistence of unstable eigenvalues for large α in more general velocity fields. In the proof, it turns out that the crucial difficulty for the analysis occurs at critical points of U rather than inflection points. A critical point of U is a point where $U' = 0$.[1]

We shall need the following assumptions:

1. All critical points are isolated. Moreover, in a neighborhood of each critical point y_c, $U''(y)(U(y) - U(y_c))$ is nonnegative.
2. If y_c is a critical point, and $U(y) = U(y_c)$, then y is also a critical point.

The first assumption is satisfied for all analytic velocity profiles. I suspect that the second assumption is not necessary, but I do not know how to avoid it in the proof. Even if the second assumption does not hold, it will be shown that growth rates of unstable modes must approach zero at an exponential rate as $\alpha \to \infty$.

2. Proof of short wave stability.

Our goal is the following result.

Theorem 2.1. *Let $U(y)$ be an analytic function defined for $y \in [0, 1]$. Moreover, assume U has the following property: If $U'(y_0) = 0$ for some $y_0 \in [0, 1]$, then $U'(y) = 0$*

*Received by the editors April 10, 2008; accepted for publication (in revised form) September 9, 2008; published electronically December 17, 2008. This research was supported by the National Science Foundation under grant DMS-0707727.

http://www.siam.org/journals/siap/69-3/72090.html

†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (renardym@math.vt.edu).

[1] This usage of the word “critical”, which is common in calculus of variations, should not be confused with a totally unrelated usage of the same word, which also occurs in hydrodynamic stability studies.

for every y with $U(y) = U(y_0)$. Then there exists α_0 such that, for $\alpha > \alpha_0$, all eigenvalues c of the Rayleigh equation (*equation* (21.17) *in* [2]),

$$(U(y) - c)(\psi''(y) - \alpha^2\psi(y)) - U''(y)\psi(y) = 0, \tag{2.1}$$

with boundary condition $\psi(0) = \psi(1) = 0$ are real.

The assumptions allow, for instance, monotone velocity profiles, profiles with a single maximum or minimum, and periodic profiles with one maximum and minimum.

For the proof, let us assume c is an eigenvalue which is not real. We can write the equation in the form

$$\psi'' - \alpha^2\psi = \frac{U''}{U - c}\psi. \tag{2.2}$$

We multiply by the conjugate of ψ and integrate, which yields

$$\int_0^1 |\psi'|^2 + \alpha^2|\psi|^2\,dy = -\int_0^1 \frac{U''(y)(U(y) - \bar{c})}{|U(y) - c|^2}|\psi|^2\,dy. \tag{2.3}$$

With $c = c_r + ic_i$, this yields the two separate equations

$$\begin{aligned}\int_0^1 |\psi'|^2 + \alpha^2|\psi|^2\,dy &= -\int_0^1 \frac{U''(y)(U(y) - c_r)}{(U(y) - c_r)^2 + c_i^2}|\psi|^2\,dy,\\ 0 &= \int_0^1 \frac{U''(y)}{(U(y) - c_r)^2 + c_i^2}|\psi|^2\,dy.\end{aligned} \tag{2.4}$$

The overall strategy of the proof is to show that, for large α, contributions to the right-hand side of the first equation in (2.4) are either negative or not large enough to balance the left-hand side. Clearly, we have

$$\left|\frac{U''(y)(U(y) - c_r)}{(U(y) - c_r)^2 + c_i^2}\right| \ll \alpha^2 \tag{2.5}$$

unless $U(y) - c_r$ is small of order α^{-2}. To focus the discussion further, we divide the points in $[0, 1]$ into the following two categories:

1. y is a regular point if $U'(y) \neq 0$.
2. y is a critical point if $U'(y) = 0$.

Since we assumed U to be analytic, all but finitely many points are regular. We shall call y a δ-regular point if the distance from the nearest critical point is at least δ.

The next lemma will be used repeatedly in what follows.

LEMMA 2.2. *Let $x_0 \in [0, 1]$, and for given $u \in L^2[0, 1]$, let*

$$Lu(x) = \frac{1}{x - x_0}\int_{x_0}^{x} u(\xi)\,d\xi. \tag{2.6}$$

Then the operator L is a bounded mapping from $L^2[0, 1]$ into itself.

This result follows immediately from Theorem 11.8 in [5], but for the sake of keeping this paper self-contained, I shall give the proof.

Let $v(x) = Lu(x)$. We have $u = ((x - x_0)v)'$, and hence $2uv - v^2 = ((x - x_0)v^2)'$ (in particular, this implies that $(x - x_0)v^2$ is continuous, since $2uv - v^2$ is integrable). Consequently,

$$\int_0^1 2uv\,dx = (1 - x_0)v(1)^2 + x_0v(0)^2 + \int_0^1 v^2\,dx, \tag{2.7}$$

which implies $\|v\| \leq 2\|u\|$. Here and in what follows, $\|\cdot\|$ refers to the norm in $L^2(0,1)$.

We shall begin with an estimate for regular points.

Lemma 2.3. *Let z be a 2δ-regular point, and let $c_r = U(\tilde{z})$ where $|z-\tilde{z}| < \delta/2$. Then there is a constant C, depending on δ but not on z and c, such that*

$$\left| \int_{z-\delta}^{z+\delta} \frac{U''(y)(U(y)-c_r)}{(U(y)-c_r)^2+c_i^2} |\psi|^2 \, dy \right| \leq C(\|\psi\|^2 + \|\psi\|^{1/2}\|\psi'\|^{3/2}). \tag{2.8}$$

If z is close to the boundary, the interval $[z-\delta,\ z+\delta]$ may not be contained in $[0,1]$. In this case, however, we can simply extend ψ by zero outside of $[0,1]$. Due to the boundary condition $\psi(0)=\psi(1)=0$, this continuation is still in H^1.

Clearly, there is a lower bound for U' on the set of all δ-regular points. Let $q(\delta)$ be this lower bound. Then we have $|U(y)-c_r| \geq q(\delta)\delta/2$ if $|y-\tilde{z}| > \delta/2$. In this range of y, we can therefore estimate the integrand by

$$\frac{2\max|U''|}{\delta q(\delta)}|\psi|^2. \tag{2.9}$$

Hence we only need to concern ourselves with

$$\int_{\tilde{z}-\delta/2}^{\tilde{z}+\delta/2} \frac{U''(y)(U(y)-c_r)}{(U(y)-c_r)^2+c_i^2}|\psi|^2 \, dy. \tag{2.10}$$

In this integral, we set

$$\psi(y) = \psi(\tilde{z}) + (y-\tilde{z})\chi(y) \tag{2.11}$$

and

$$|\psi(y)|^2 = |\psi(\tilde{z})|^2 + (y-\tilde{z})(\psi(y)\bar{\chi}(y) + \chi(y)\bar{\psi}(\tilde{z})). \tag{2.12}$$

According to Lemma 2.2 above, we have $\|\chi\| \leq 2\|\psi'\|$, and from the trace theorem (see (3.18) in [5]) we have $|\psi(\tilde{z})|^2 \leq C(\|\psi\|^2 + \|\psi\|\|\psi'\|)$ for some constant C. Moreover,

$$\frac{U''(y)(U(y)-c_r)(y-\tilde{z})}{(U(y)-c_r)^2+c_i^2} \tag{2.13}$$

is bounded. Consequently, we find

$$\begin{aligned} &\left| \int_{\tilde{z}-\delta/2}^{\tilde{z}+\delta/2} \frac{U''(y)(U(y)-c_r)(y-\tilde{z})}{(U(y)-c_r)^2+c_i^2}(\psi(y)\bar{\chi}(y) + \chi(y)\bar{\psi}(\tilde{z}))\, dy \right| \\ &\quad \leq C(\delta)(\|\psi\|\|\chi\| + |\psi(\tilde{z})|\|\chi\| \leq C(\delta)(\|\psi\|^2 + \|\psi'\|^{3/2}\|\psi\|^{1/2}). \end{aligned} \tag{2.14}$$

It remains to estimate

$$|\psi(\tilde{z})|^2 \left| \int_{\tilde{z}-\delta/2}^{\tilde{z}+\delta/2} \frac{U''(y)(U(y)-c_r)}{(U(y)-c_r)^2+c_i^2}\, dy \right|. \tag{2.15}$$

In this last integral, we substitute $U(y)$ as a new variable to obtain the new integral

$$\int_{U(\tilde{z}-\delta/2)}^{U(\tilde{z}+\delta/2)} \frac{U''(y(U))(U-c_r)}{U'(y(U))[(U-c_r)^2+c_i^2]}\, dU. \tag{2.16}$$

The interval of integration contains the symmetric interval $|U - c_r| < \delta q(\delta)/2$, and outside this interval, the integrand is bounded by a constant $C(\delta)$. Next, we write

$$\frac{U''(y(U))}{U'(y(U))} = \frac{U''(\tilde{z})}{U'(\tilde{z})} + (U - c_r)S(U), \tag{2.17}$$

where $S(U)$ is a continuous function. By symmetry, we then find

$$\int_{c_r-\delta q(\delta)/2}^{c_r+\delta q(\delta)/2} \frac{U''(y(U))(U - c_r)}{U'(y(U))[(U - c_r)^2 + c_i^2]}\, dU = \int_{c_r-\delta q(\delta)/2}^{c_r+\delta q(\delta)/2} \frac{(U - c_r)^2 S(U)}{(U - c_r)^2 + c_i^2}\, dU. \tag{2.18}$$

The integrand in the latter integral is bounded by a constant. This completes the proof of Lemma 2.3.

We next consider the neighborhood of a critical point.

LEMMA 2.4. *There exist $\epsilon > 0$ and $K > 0$ such that, if y_0 is in an ϵ-neighborhood of a critical point y_c, and $c_r = U(y_0)$, then*

$$\int_0^1 \frac{U''(y)(U(y) - c_r)}{(U(y) - c_r)^2 + c_i^2} |\psi|^2\, dy \geq -K\|\psi\|^2. \tag{2.19}$$

We exploit the second equation of (2.4) to obtain that

$$\int_0^1 \frac{U''(y)(U(y) - c_r)}{(U(y) - c_r)^2 + c_i^2} |\psi|^2\, dy = \int_0^1 \frac{U''(y)(U(y) - U(y_c))}{(U(y) - c_r)^2 + c_i^2} |\psi|^2\, dy. \tag{2.20}$$

All values of y where $U(y) = U(y_c)$ are critical points, and there is a finite number of these. Each of them has a neighborhood on which $U''(y)(U(y) - U(y_c))$ is nonnegative. If we choose ϵ small enough, then $|U(y) - c_r|$ has a positive lower bound outside these neighborhoods. The lemma follows.

To prove the theorem, we need to bound the right-hand side in the first equation of (2.4). Lemma 2.3 gives a bound of the form

$$C(\|\psi\|^2 + \|\psi\|^{1/2}\|\psi'\|^{3/2}) \tag{2.21}$$

in the neighborhood of regular points. Lemma 2.4 gives an upper bound of the form $K\|\psi\|^2$ in the neighborhood of critical points. By combining the two, we find that the right-hand side in the first equation of (2.4) cannot balance the left-hand side if α is large and ψ is nontrivial, which is the desired result.

We note that the assumption of analyticity was used only to ensure that the number of critical points is finite and that $(U(y) - U(y_c))U''(y)$ is nonnegative in a neighborhood of each critical point.

Without the assumption that U assumes its critical values only at critical points, we can still prove that the growth rate of unstable modes must tend to zero at an exponential rate as $\alpha \to \infty$.

THEOREM 2.5. *Let U be any analytic function on $[0, 1]$. Then there exists a function $s(\alpha)$, with $s(\alpha) \to 0$ for $\alpha \to \infty$, such that, if c is a nonreal eigenvalue of the Rayleigh equation, and α is sufficiently large, then there exists a critical point y_c with $|c - U(y_c)| \leq s(\alpha)$. Moreover, there exist constants C and k such that $|\alpha c_i| \leq C|c - U(y_c)| \exp(-k|\alpha|)$.*

The first statement, that c_r must be close to a critical value of U when α is large, follows from the proof of the previous theorem. Now let ρ be a nonnegative function which vanishes in an ϵ-neighborhood of the critical points but is equal to 1 at distance more than 2ϵ from the critical points. We multiply the Rayleigh equation by $\rho\bar{\psi}$ and integrate. The result is

$$\int_0^1 \rho(|\psi'|^2 + \alpha^2|\psi|^2)\,dy = -\int_0^1 \frac{U''(y)(U(y)-\bar{c})}{|U(y)-c|^2}\rho|\psi|^2\,dy - \int_0^1 \rho'\psi'\bar{\psi}\,dy. \tag{2.22}$$

With the help of Lemma 2.3, this yields the estimate

$$\int_0^1 \rho(|\psi'|^2 + \alpha^2|\psi|^2)\,dy \le C(\|\psi\|^2 + \|\psi\|^{1/2}\|\psi'\|^{3/2}). \tag{2.23}$$

Now let y_c be a critical point, and let χ be a smooth function which is equal to 1 in an ϵ-neighborhood of y_c and has support in a 2ϵ-neighborhood of y_c. We can choose ϵ such that $U''(U - U(y_c))$ is nonnegative on the support of χ. We shall also assume that $c_r = U(y_1)$, where y_1 is within $\epsilon/2$ of y_c. We multiply the Rayleigh equation by $\chi\bar{\psi}$ and integrate. The result is

$$\int_0^1 \chi(|\psi'|^2 + \alpha^2|\psi|^2)\,dy = -\int_0^1 \frac{U''(y)(U(y)-\bar{c})}{|U(y)-c|^2}\chi|\psi|^2\,dy - \int_0^1 \chi'\psi'\bar{\psi}\,dy. \tag{2.24}$$

Taking the imaginary part of this identity, we conclude that

$$|c_i|\left|\int_0^1 \frac{U''}{|U-c|^2}\chi|\psi|^2\,dy\right| \le \left|\int_0^1 \chi'\psi'\bar{\psi}\,dy\right|. \tag{2.25}$$

For large α, the solutions of the Rayleigh equation have exponential asymptotics as long as we stay away from points where $U - c$ is small (see Theorem 26.3 in [6]). Since the support of χ' is separated from the points where $|U - c|$ is small, we can conclude that there is a bound of the form

$$\left|\int_0^1 \chi'\psi'\bar{\psi}\,dy\right| \le C\exp(-k(\epsilon)|\alpha|)\|\psi\|^2. \tag{2.26}$$

Next, we consider the real part of (2.24), which we write in the form

$$\begin{aligned}&\int_0^1 \chi\left(|\psi'|^2 + \alpha^2|\psi|^2 + \frac{U''(U-U(y_c))}{|U-c|^2}|\psi|^2\right)\,dy\\ &= \int_0^1 \frac{U''(c_r - U(y_c))}{|U-c|^2}\chi|\psi|^2 - \chi'\,\mathrm{Re}\,(\psi'\bar{\psi})\,dy.\end{aligned} \tag{2.27}$$

We can now use (2.25) and (2.26) to bound the right-hand side by

$$C\left(1 + \frac{|c_r - U(y_c)|}{|c_i|}\right)\exp(-k(\epsilon)|\alpha|)\|\psi\|\|\psi'\|. \tag{2.28}$$

By combining this result with (2.23), we conclude that we have a bound of the form

$$\|\psi'\|^2 + \alpha^2\|\psi\|^2 \le \frac{|c_r - U(y_c)|\exp(-k(\epsilon)|\alpha|)}{|c_i|}\|\psi\|\|\psi'\|. \tag{2.29}$$

This is not possible if $|\alpha c_i| \gg |c_r - U(y_c)|\exp(-k(\epsilon)|\alpha|)$.

REFERENCES

[1] P. G. Drazin and L. N. Howard, *Hydrodynamic stability of parallel flow of inviscid fluid*, Adv. Appl. Mech., 9 (1966), pp. 1–89.

[2] P. G. Drazin and W. H. Reid, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1981.

[3] S. Friedlander and A. Lipton-Lifschitz, *Localized instabilities in fluids*, in Handbook of Mathematical Fluid Dynamics 2, S. Friedlander and D. Serre, eds., North–Holland, Amsterdam, 2003, pp. 289–354.

[4] L. N. Howard, *The number of unstable modes in hydrodynamic stability problems*, J. Mécanique, 3 (1964), pp. 433–443.

[5] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer, Berlin, 1972.

[6] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Krieger, Huntington, 1976.

SIAM J. APPL. MATH.
Vol. 69, No. 3, pp. 769–786

STEADY AND INTERMITTENT SLIPPING IN A MODEL OF LANDSLIDE MOTION REGULATED BY PORE-PRESSURE FEEDBACK[*]

DAVID G. SCHAEFFER[†] AND RICHARD M. IVERSON[‡]

Abstract. This paper studies a parsimonious model of landslide motion, which consists of the one-dimensional diffusion equation (for pore pressure) coupled through a boundary condition to a first-order ODE (Newton's second law). Velocity weakening of sliding friction gives rise to nonlinearity in the model. Analysis shows that solutions of the model equations exhibit a subcritical Hopf bifurcation in which stable, steady sliding can transition to cyclical, stick-slip motion. Numerical computations confirm the analytical predictions of the parameter values at which bifurcation occurs. The existence of stick-slip behavior in part of the parameter space is particularly noteworthy because, *unlike stick-slip behavior in classical models*, here it arises in the absence of a reversible (elastic) driving force. Instead, the driving force is static (gravitational), mediated by the effects of pore-pressure diffusion on frictional resistance.

Key words. Hopf bifurcation, landslide, pore pressure, stick-slip

AMS subject classifications. 74H60, 74F10, 74C99, 37G15

DOI. 10.1137/07070704X

1. Introduction.

Landslides exhibit a great diversity of movement styles and rates, including steady creeping slip, intermittent rapid slip, and catastrophic avalanching. Recently Iverson [4, 5] introduced a new theoretical model that in numerical simulations exhibits all these behaviors as a consequence of pore-pressure feedback. Most intriguing is the transition between steady and intermittent slip. In this paper we analyze this transition as a bifurcation problem.

As sketched in Figure 2.1, consider a block of porous soil on a rigid planar slope that is inclined at an angle θ. If there is no liquid in the pores of the soil, then friction can support its weight at rest provided μ_0, the coefficient of static friction, is greater than $\tan\theta$. If, however, water pressure acting on the base of the block is sufficiently large, the block will begin to slide. Suppose that it slides rigidly except for a zone of intense shearing at its base. If the soil in the basal shear zone is compacted, the governing equations admit a solution with steady sliding, due to the following sequence of physical effects:

> As the basal zone shears, it dilates; this expansion creates new pore space, thereby reducing the fluid pressure in the expanding pores; in consequence, the normal traction on the soil matrix is increased; and increased friction between the soil and the base can balance the driving and resisting forces, leading to steady creep as water pressure is restored by steady diffusion from the overlying slide block.

If friction is rate-independent, this steady creeping motion is stable. However, even a small amount of rate softening in the friction law is sufficient to destabilize steady

[*] Received by the editors October 31, 2007; accepted for publication (in revised form) August 5, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/siap/69-3/70704.html

[†] Math Dept., Duke University, Durham, NC 27708 (dgs@math.duke.edu). The research of this author was supported under NSF grant DMS-0244492.

[‡] U.S. Geological Survey, Cascades Volcano Observatory, Vancouver, WA 98686 (riverson@usgs.gov).

motion through a Hopf bifurcation. This bifurcation is the primary focus of the present paper.

Hopf bifurcation explains the origin of oscillatory behavior in the system. Intermittent behavior—brief periods of rapid slipping alternating with comparatively long periods with no slipping—arises from the singular behavior of friction at zero velocity. Specifically, the resisting frictional force jumps to a static value dictated by the ambient pore pressure. The block then remains stationary while diffusion brings pore pressure back up to a level where friction can no longer balance gravity.

The outline of this paper is as follows. In section 2 we introduce the equations of Iverson's model, nondimensionalize them, and linearize them around the steady-state solution. Mathematically, provided the block velocity is positive, this model may be described as a parabolic PDE for the pore pressure coupled through a boundary condition to an ODE for the block velocity. In section 3 the linearized equations are solved by separation of variables, leading to a trancendental equation for eigenvalues. In section 4 we extract the condition for bifurcation by analyzing the eigenvalue equation. In section 5 we summarize the results of supporting computations, which agree well with the theoretical predictions. In section 6 we present a concluding discussion regarding our findings. Finally, in an appendix we provide a mathematical proof omitted from the main text.

Mathematically, this model is interesting in that time-periodic behavior appears in a problem governed by the (scalar) diffusion equation (of course, coupled to an ODE through a boundary condition). Physically, the model is important because it provides a parsimonious mathematical description of diverse landslide behavior that has not been rigorously analyzed until now.

2. Governing equations.

2.1. Dimensional formulation. As indicated in Figure 2.1, consider a block of soil of porosity ϕ, height H_s, and density ρ_s, which is defined as the mass of solid grains per unit total volume (i.e., the porosity is factored into ρ_s). The block is saturated with pore water of density ρ_w to a height H_w that does not change with time.[1]

Suppose this system is supported by a planar slope inclined at an angle θ. Using coordinates aligned with the slope, we describe this system by the pore pressure $p(y,t)$, the traction $\tau_x(t), \tau_y(t)$ exerted by the supporting plane on the solid matrix (effective stress at the base), and, assuming the block slides as a rigid unit over a shearing basal zone, the velocity $v_x(t), v_y(t)$ of the block. The motion is assumed to be one-dimensional in that all variables are independent of the tangential coordinate x, but it is two-dimensional in that the block is allowed to move in the normal direction as well as the tangential—indeed, dilatancy of the shearing basal zone requires this.

Let us decompose the total pore pressure into a hydrostatic component plus the excess pore pressure associated with dilation,

$$p_{\text{tot}}(y,t) = p_{\text{hydro}}(y) + p_{\text{ex}}(y,t),$$

[1] The assumption that H_w is constant is not strictly satisfied owing to small fluxes of water to and from the basal shear zone. Changes in H_w can be estimated from simple mass-balance considerations. On this basis we estimate that such changes are less than 1%. We exclude such changes from our model, however, not only for the sake of simplicity, but also because rigorous assessment of changes in water-table height requires consideration of hysteretic, nonlinear processes associated with variably saturated groundwater flow [2].

Incidentally, the model can easily be modified to allow for flux of water into the soil beneath the shear zone as well as flux into the soil above the shear zone. Computations with these options did not differ qualitatively from those reported here.

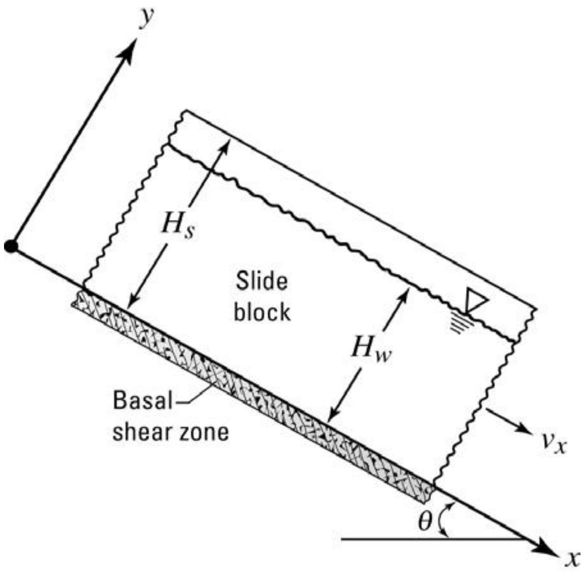


FIG. 2.1. *Schematic of landslide block.*

where

$$p_{\text{hydro}}(y) = \rho_w g \cos\theta(H_w - y). \tag{2.1}$$

Following arguments summarized by Iverson [5], we suppose the excess pore pressure evolves diffusively,

$$\partial_t p_{\text{ex}} = D\partial_{yy} p_{\text{ex}}, \qquad 0 < y < H_w, \tag{2.2}$$

with boundary conditions

$$\begin{array}{lrcl} \text{(a)} & p_{\text{ex}}(H_w, t) & = & 0, \\ \text{(b)} & \partial_y p_{\text{ex}}(0, t) & = & (\rho_w g/K) v_y(t), \end{array} \tag{2.3}$$

where D is the saturated hydraulic diffusivity, K is the saturated hydraulic conductivity below the water table, and g is the acceleration of gravity. Typical values for these parameters, and for others below, are given in Table 2.1. Equation (2.3)(b) follows from Darcy's law [2]: the excess-pressure gradient at the boundary of the shear zone is proportional to the fluid flux through the boundary that is needed to fill the volume vacated through dilatancy.

In Iverson's model, the behavior of the basal zone is characterized simply by two constitutive equations: (i) dilatancy,

$$v_y = \psi v_x, \tag{2.4}$$

where ψ is the angle of dilatancy, and (ii) friction,

$$\begin{array}{lccccccc} \text{either} & v_x & = & 0 & \text{and} & |\tau_x| & < & \mu_0 \tau_y, \\ \text{or} & \tau_x & = & -\mu(v_x)\tau_y & \text{and} & v_x & \geq & 0, \end{array} \tag{2.5}$$

where $\mu(v_x)$ is a rate-softening coefficient of friction,

$$\mu(v_x) = \mu_0[1 - a \sinh^{-1}(v_x/2v_{\text{ref}})], \tag{2.6}$$

TABLE 2.1
Parameter values in the model.

Parameter	Definition	Units	Plausible values	Values used in computations
a	Rate-dependence coefficient in friction rule	—	0 — 0.05	0.02, 0.04
D	Hydraulic diffusivity of soil	m^2/s	10^{-8} — 1	10^{-8} — 3×10^{-1}
g	Acceleration of gravity	$\mathrm{m/s}^2$	9.8	9.8
H_s	Thickness of soil block	m	0.1 — 100	0.65
H_w	Height of water table	m	0 — 100	0.3701
K	Hydraulic conductivity of soil	m/s	10^{-11} — 1	2×10^{-9} — 2×10^{-1}
v_{ref}	Reference slip rate in friction rule	m/s	10^{-4} — 1	7×10^{-4} — 4×10^{-1}
θ	Slope angle	radians	0.1 — 0.8	0.5411
μ_0	Static friction coefficient of soil	—	0.2 — 1.2	0.8693
ρ_w	Mass density of pore water	$\mathrm{kg/m}^3$	900 — 1100	1000
ρ_s	Bulk density of dry soil	$\mathrm{kg/m}^3$	1000 — 2200	1600
ϕ	Porosity of soil block	—	0.2 — 0.6	0.4
ψ	Dilatancy angle of shearing soil	radians	-0.2 — 0.2	0.1047

v_{ref} being a reference velocity. Regarding dilatancy, we take ψ to be constant. Under continued shearing, ψ will in fact converge to zero as the soil tends to a critical state. However, as shown in the landslide experiments of [6], the evolution to critical state occurs only when displacement magnitudes greatly exceed those appearing during the processes considered here. Equation (2.6) is a three-parameter representation of rate-weakening friction that we use instead of a more complicated, rate-and-state friction model, which typically employs six parameters. Because of its simplicity, we are able to explore the full relevant parameter space. However, investigation of the model with a rate-and-state friction law remains a task for future research.

Finally, the system is completed by Newton's equations of motion for the block. For this system, the mass per unit area of slope is $\rho_s H_s + \phi\rho_w H_w$. Thus, for the tangential component of the motion we have

$$(\rho_s H_s + \phi\rho_w H_w)\partial_t v_x = (\rho_s H_s + \phi\rho_w H_w) g \sin\theta + \tau_x. \tag{2.7}$$

Regarding the normal component, we are assuming the dilatancy ψ is small, and therefore, in light of (2.4), we may neglect acceleration in the y-direction; thus Newton's equation reduces to force balance. According to Terzaghi's effective-stress principle (e.g., see [2]), the effective normal traction exerted by the slope on the solid matrix is the total stress reduced by the pore pressure at the base; in symbols,

$$\tau_y = (\rho_s H_s + \phi\rho_w H_w) g \cos\theta - p_{\mathrm{tot}}(0,t).$$

On substitution of (2.1), we obtain our last equation,

$$\tau_y = [\rho_s H_s - (1-\phi)\rho_w H_w] g \cos\theta - p_{\mathrm{ex}}(0,t). \tag{2.8}$$

This formulation differs from that of Iverson [5] in two main respects:

- Most importantly, here we allow for rate-softening friction.
- Our assumptions on the imposed pore pressure are more restrictive: specifically, in the notation of (8) of Iverson [5], we assume that $\beta = \cos\theta$ and $W = 0$. Physically, these assumptions imply that there is no flux of groundwater normal to the water table, except for the flux caused by shear-zone dilation.

Provided

$$(\rho_s H_s + \phi\rho_w H_w)\sin\theta \le \mu_0\left[\rho_s H_s - (1-\phi)\rho_w H_w\right]\cos\theta, \tag{2.9}$$

the above equations have a solution with $p_{\mathrm{ex}} = 0$ and $v = 0$: i.e., friction is sufficient to resist the pull of gravity. We study the case where (2.9) is violated.

2.2. Nondimensionalization. We nondimensionalize (2.2)–(2.8) by defining

$$t = \frac{t^{(\mathrm{dim})}}{H_w^2/D}, \quad y = \frac{y^{(\mathrm{dim})}}{H_w}, \quad v = \frac{v_x^{(\mathrm{dim})}}{K}, \quad p = \frac{p_{\mathrm{ex}}^{(\mathrm{dim})}}{\rho_w g H_w}, \quad \tau = \frac{\tau^{(\mathrm{dim})}}{\rho_s g H_s}, \tag{2.10}$$

where the superscript *dim* indicates the dimensional version of a variable. We will eliminate $v_y^{(\mathrm{dim})}$ from the equations, so we do not define a scaled version of this variable; however, we nondimensionalize both components of τ. As in Iverson [5], we have used the diffusive time scale to nondimensionalize t; however, our nondimensionalization of v differs from that of [5]. We also define two dimensionless constants that will appear in the nondimensionalized equations below,

$$\varepsilon = \frac{K/g}{H_w^2/D} \qquad \text{and} \qquad M = \frac{\rho_w H_w}{\rho_s H_s}. \tag{2.11}$$

The first, which according to Table 2.1 is very small, is the ratio of the acceleration time scale to the diffusive time scale; the second is ϕ^{-1} times the ratio of fluid mass to solid mass.

The evolution of the nondimensionalized pressure is governed by

$$\begin{array}{llcll} \text{(a)} & \partial_t p & = & \partial_{yy} p, & 0 < y < 1, \\ \text{(b)} & p(1,t) & = & 0, & \\ \text{(c)} & \partial_y p(0,t) & = & \psi v(t). & \end{array} \tag{2.12}$$

In nondimensional variables, the friction relation (2.5) does not change, except that the rate-softening coefficient in (2.6) must be rescaled to give

$$\mu(v) = \mu_0\left[1 - a\sinh^{-1}\left(\frac{K}{2v_{\mathrm{ref}}}v\right)\right]. \tag{2.13}$$

Newton's equations for the motion of the block scale to

$$\begin{array}{llcl} \text{(a)} & \varepsilon\,\partial_t v & = & \sin\theta - (1+\phi M)^{-1}\tau_x, \\ \text{(b)} & \tau_y & = & [1-(1-\phi)M]\cos\theta - Mp_{\mathrm{ex}}(0,t). \end{array} \tag{2.14}$$

In nondimensional variables, the no-motion condition (2.9) may be rewritten as

$$\tan\theta \le A_1\mu_0, \tag{2.15}$$

where A_1 is the first of two mass ratios defined in (2.17) below. If (2.15) is violated and hence $v > 0$, then (2.5) and (2.14)(b) may be combined to solve for τ_x. On substitution into (2.14)(a), we obtain

$$\varepsilon\,\partial_t v = \sin\theta - \mu(v)\left(A_1\cos\theta - A_2 p(0,t)\right), \tag{2.16}$$

where

$$A_1 = \frac{1-(1-\phi)M}{1+\phi M}, \qquad A_2 = \frac{M}{1+\phi M}. \tag{2.17}$$

As long as $v > 0$, the motion is described by (2.12), (2.16).

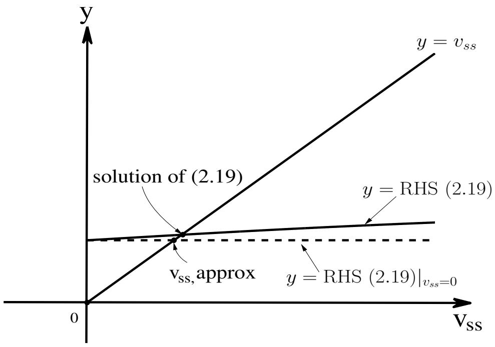


FIG. 2.2. *Graphical solution of* (2.19).

2.3. A steady solution and linearization of the equations. Henceforth we assume that

$$(2.15) \text{ is violated and } \psi > 0.$$

Let us look for a steady-state solution $p_{\mathrm{ss}}(y), v_{\mathrm{ss}}$ of the equations of motion. It follows from (2.12) that

$$p_{\mathrm{ss}}(y) = \psi v_{\mathrm{ss}}(y-1), \tag{2.18}$$

and this relation may be substituted into (2.16) to obtain an implicit equation for v_{ss},

$$v_{\mathrm{ss}} = (A_2\psi)^{-1}\left[\frac{\sin\theta}{\mu(v_{\mathrm{ss}})} - A_1\cos\theta\right]. \tag{2.19}$$

If the friction coefficient is independent of velocity, then this equation is in fact a formula for v_{ss}. Even with rate-softening friction, for parameter values such as in Table 2.1, the right-hand side (RHS) of (2.19) is a slowly varying function of v_{ss} (see Figure 2.2). For example, defining

$$v_{\mathrm{ss,approx}} = (A_2\psi)^{-1}\left[\frac{\sin\theta}{\mu_0} - A_1\cos\theta\right] \tag{2.20}$$

as the solution of (2.19) in the rate-independent case, we may see that

$$\frac{\mu(v_{\mathrm{ss,approx}})}{\mu_0} = 1 - \mathcal{O}\left(\frac{a}{\psi}\frac{K}{v_{\mathrm{ref}}}\right),$$

and, unless ψ is extremely small, we have

$$\frac{a}{\psi}\frac{K}{v_{\mathrm{ref}}} \ll 1. \tag{2.21}$$

Thus, assuming $\psi > 0$ is not too small, we conclude that (2.19) has a unique solution close to $v_{\mathrm{ss,approx}}$.

Incidentally, there is a second solution of (2.19) at very large values of v_{ss}. To see this, observe from (2.13) that $\mu(v)$ vanishes when $v \approx (v_{\mathrm{ref}}/K)e^{1/a}$. Since $\mu(v_{\mathrm{ss}})$ occurs in the denominator, the RHS of (2.19) blows up as v tends to $(v_{\text{ref}}/K)e^{1/a}$, giving rise to a second intersection with the linear function on the left-hand side (LHS). This second steady-state solution of (2.19), for which the velocity is very large indeed, is always unstable, but it can be involved in the steady-state bifurcation of (2.12), (2.16) at extreme parameter values. For clarity, we shall refer to the first solution, given approximately by (2.20), as the *physical* steady state. Referring to the figure and comparing the derivatives of both sides of (2.19), we conclude that at the physical steady state

$$\frac{\sin\theta\,|\mu'(v_{\text{ss}})|}{A_2\psi\,\mu^2(v_{\text{ss}})} < 1. \tag{2.22}$$

To linearize (2.12), (2.16) near the (physical) steady-state solution, we define incremental variables $\overline{p}, \overline{v}$ by

$$p(y,t) = p_{\text{ss}}(y) + \overline{p}(y,t), \qquad v(t) = v_{\text{ss}} + \overline{v}(t).$$

Equations (2.12) are already linear, so we find trivially that

$$\begin{array}{lrcll} \text{(a)} & \partial_t\overline{p} & = & \partial_{yy}\overline{p}, & 0 < y < 1, \\ \text{(b)} & \overline{p}(1,t) & = & 0, & \\ \text{(c)} & \partial_y\overline{p}(0,t) & = & \psi\overline{v}(t), & \end{array} \tag{2.23}$$

and on linearizing (2.16) and simplifying using (2.19) we obtain

$$\varepsilon\partial_t\overline{v} = B_1\overline{p}(0,t) + B_2,\overline{v}, \tag{2.24}$$

where

$$B_1 = A_2\,\mu(v_{\text{ss}}) \quad \text{and} \quad B_2 = \sin\theta\,|\mu'(v_{\text{ss}})|/\mu(v_{\text{ss}}). \tag{2.25}$$

Since $\mu'(v_{\text{ss}}) < 0$, we have used the absolute value to emphasize that $B_2 > 0$.

3. Derivation of the eigenvalue equation. We seek a solution of the linearized equations (2.23), (2.24) with exponential time dependence $e^{-\lambda t}$ (note the minus sign). Using separation of variables, we derive from (2.23)(a),(b) that

$$\overline{p}(y,t) = P\sin[\sqrt{\lambda}(1-y)]e^{-\lambda t}, \qquad \overline{v}(t) = Ve^{-\lambda t},$$

where P and V are constants. Substitution of these formulas into (2.23)(c), (2.24) yields the 2×2 homogeneous linear system

$$\begin{bmatrix} \sqrt{\lambda}\cos\sqrt{\lambda} & \psi \\ B_1\sin\sqrt{\lambda} & \varepsilon\lambda + B_2 \end{bmatrix} \begin{bmatrix} P \\ V \end{bmatrix} = 0. \tag{3.1}$$

This system has a nonzero solution if and only if the determinant of the coefficient matrix vanishes, which leads to the trancendental equation for the decay rate λ,

$$\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} = \varepsilon C_1\lambda + C_2, \tag{3.2}$$

where

$$C_1 = 1/B_1\psi, \qquad C_2 = B_2/B_1\psi. \tag{3.3}$$

Recalling (2.25), (2.22), we see that

$$C_2 < 1. \tag{3.4}$$

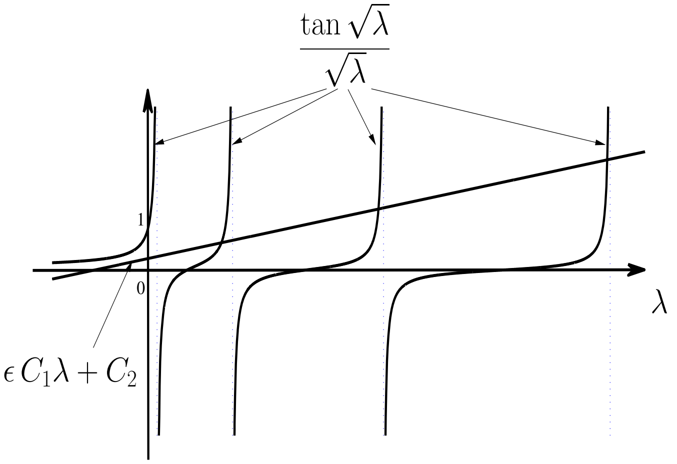


FIG. 4.1. *Graphical determination of the real solutions of* (3.2). *In the figure shown,* $\varepsilon C_1 = 0.02$ *and* $C_2 = 0.3$. *As indicated by* (4.9), *the two complex solutions of* (3.2) *lie in the unstable half-plane,* $\{\mathrm{Re}\,\lambda < 0\}$.

4. Analysis of the eigenvalue equation.

4.1. Introduction. As illustrated in Figure 4.1, (3.2) has an infinite sequence of positive roots. Since these eigenvalues are all in the stable half plane, they do not require further attention. It is not obvious, but (3.2) has two other, possibly complex, roots, which are the focus of the present section.

As a function of a complex variable, $(\tan\sqrt{\lambda})/\sqrt{\lambda}$ is a meromorphic function: i.e., apart from a sequence of poles on the positive real axis, it is single-valued and analytic in the entire plane. Although neither the numerator nor the denominator of this expression is single-valued, the quotient avoids this difficulty. Of course, we use the same branch of $\sqrt{\lambda}$ in the numerator and the denominator so that

$$\lim_{\lambda\to 0}\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} = 1.$$

To be specific, let us choose the branch

$$\sqrt{\lambda} = |\lambda|^{1/2}\, e^{i(\arg\lambda)/2}, \tag{4.1}$$

where $\arg\lambda$ satisfies

$$0 \le \arg\lambda < 2\pi.$$

The analysis of the complex roots of (3.2) is based on the simple behavior of $\tan\sqrt{\lambda}$ away from the positive real axis, as articulated in the following proposition.

PROPOSITION 4.1. *Let* Λ *be a wedge in* $\mathbb{C}$ *excluding the positive real axis, say,*

$$\Lambda = \{\lambda \in \mathbb{C} : \delta < \arg\lambda < 2\pi - \delta\}, \tag{4.2}$$

where $0 < \delta < \pi/2$. *Then, as* $|\lambda| \to \infty$ *in* Λ,

$$\tan\sqrt{\lambda} = i + \mathcal{O}(e^{-\delta\sqrt{|\lambda|}/2}). \tag{4.3}$$

Proof. By manipulating the definition of $\tan z$, we deduce that

$$\tan z = i\,\frac{1-e^{2iz}}{1+e^{2iz}}. \tag{4.4}$$

Now $|e^{2iz}| = e^{-2\operatorname{Im} z}$, so taking $z=\sqrt{\lambda}$ we see that

$$\tan\sqrt{\lambda} = i + \mathcal{O}(e^{-2\operatorname{Im}\sqrt{\lambda}}).$$

By (4.1), $\operatorname{Im}\sqrt{\lambda} = |\lambda|^{1/2}\sin(\arg\lambda/2)$. To complete the proof, we estimate $\arg\lambda$ with (4.2) and use the fact that $\sin(\delta/2)\geq\delta/4$. □

4.2. The rate-independent case ($C_2 = 0$). When friction is independent of velocity, the coefficient C_2 in (3.2) vanishes. In the appendix we prove that in this case (3.2) has no zeros in the left half plane: i.e., the steady solution is stable. Since the proof sheds little light on the Hopf bifurcation, we do not include it here. It is instructive, however, to locate the two complex eigenvalues.

Suppose $C_2 = 0$. By (4.3), for large $|\lambda|$ away from the positive real axis, (3.2) may be rewritten, approximately, as

$$e^{i[\pi-\arg\lambda]/2} = \varepsilon C_1|\lambda|^{3/2}e^{i\arg\lambda}.$$

Equating magnitudes we find that $|\lambda| = (\varepsilon C_1)^{-2/3}$, and then equating arguments we find the two approximate roots of (3.2):

$$\lambda = e^{i\pi/3}(\varepsilon C_1)^{-2/3},\quad e^{5i\pi/3}(\varepsilon C_1)^{-2/3}. \tag{4.5}$$

By (4.3) the error in this estimate is exponentially small in ε. Since $\operatorname{Re}\lambda \gg 1$, the associated eigenfunctions decay rapidly in time.

4.3. The rate-dependent case: Steady-state bifurcation. If C_2 assumes positive values, the complex[2] eigenvalues (4.5) can cause the linearized equations (2.23), (2.24) to lose stability if they cross into the left half plane. As we shall see in the next subsection, for physical parameter values, (2.23), (2.24) lose stability through a Hopf bifurcation: i.e., the complex eigenvalues cross the imaginary axis as a pair of complex conjugates. However, for mathematical completeness, we also ask when real solutions of (3.2) cross the imaginary axis. Indeed, one may see by inspection that $\lambda = 0$ is a root of (3.2) iff $C_2 = 1$. If one forces C_2 to its limiting value unity (cf. (3.4)), then the two solutions of (2.19) merge and annihilate one another at a steady-state bifurcation of limit-point type [3].

4.4. The rate-dependent case: Hopf bifurcation.

(a) *Main calculations.* Figure 4.2 shows a curve Γ in the $(\varepsilon C_1, C_2)$-plane that separates the infinite strip

$$\{(\varepsilon C_1, C_2)\ :\ 0<\varepsilon C_1<\infty,\ 0<C_2<1\}$$

into two regions in which (2.23), (2.24) are stable or unstable. This curve has the parametric representation

$$\varepsilon C_1 = \frac{2}{\mu^3}\,\frac{1-2e^{-\mu}\sin\mu - e^{-2\mu}}{1+2e^{-\mu}\cos\mu + e^{-2\mu}}, \tag{4.6}$$

[2] We shall refer to these roots of (3.2) as *complex* eigenvalues even though, for extreme parameter values, they may actually become real.

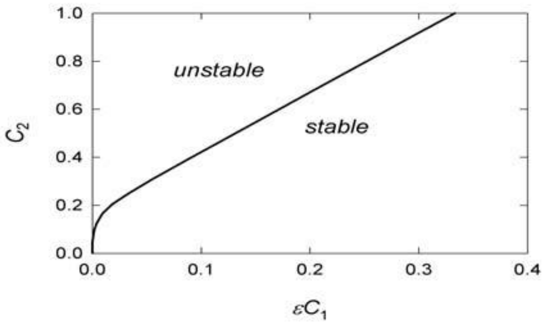


FIG. 4.2. *Graph of the Hopf bifurcation curve* Γ *defined parametrically by* (4.6), (4.7).

$$C_2 = \frac{1}{\mu}\,\frac{1 + 2e^{-\mu}\sin\mu - e^{-2\mu}}{1 + 2e^{-\mu}\cos\mu + e^{-2\mu}}, \tag{4.7}$$

where $0 \le \mu < \infty$. In the following proposition we show that (2.23), (2.24) undergo Hopf bifurcation when C_1, C_2 lies on Γ.

PROPOSITION 4.2. *If* $\varepsilon C_1, C_2$ *lie on* Γ, *then the complex eigenvalues of* (3.2) *are pure imaginary.*

Proof. Suppose (3.2) has a root on the positive imaginary axis, say, at $\lambda = i\mu^2/2$, where $\mu \ge 0$. Then $\sqrt{\lambda} = (1+i)\mu/2$. Equating real and imaginary parts of (3.2), we conclude that

$$\varepsilon C_1 = \frac{2}{\mu^2}\,\mathrm{Im}\left\{\frac{\tan[(1+i)\mu/2]}{(1+i)\mu/2}\right\}, \qquad \varepsilon C_2 = \mathrm{Re}\left\{\frac{\tan[(1+i)\mu/2]}{(1+i)\mu/2}\right\}. \tag{4.8}$$

We claim that

$$\tan[(1+i)\mu/2] = \frac{2e^{-\mu} + i(1 - e^{-2\mu})}{1 + 2e^{-\mu}\cos\mu + e^{-2\mu}},$$

which may be proved by recalling (4.4), multiplying and dividing by the complex conjugate of the denominator, and simplifying. Equations (4.6), (4.7) result on multiplying by $[(1+i)\mu/2]^{-1} = (1-i)/\mu$ and substituting into (4.8). □

By examining the Taylor series expansions of the numerators in (4.6), (4.7), we see that

$$\lim_{\mu\to 0} \varepsilon C_1 = 1/3, \qquad \lim_{\mu\to 0} C_2 = 1,$$

which is behavior that may be seen in Figure 4.2. At the other extreme, $\mu \gg 1$, or equivalently $\varepsilon C_1 \ll 1$, the exponentials in (4.6), (4.7) may be neglected, so that it is possible to eliminate μ and obtain a relation between the C's that characterizes Hopf bifurcation:

$$C_2 = \left(\frac{\varepsilon C_1}{2}\right)^{1/3}, \tag{4.9}$$

both sides of the equation being small. The proof of the proposition shows that at parameter values given by (4.6), (4.7), the complex eigenvalues of (2.23), (2.24) are $\pm i\mu^2/2$, and in the asymptotic range the complex eigenvalues are approximately $\pm iC_2^{-2}/2$. In particular, the bifurcating periodic solutions have periods approximately equal to

$$\frac{2\pi}{|\lambda|} \approx 4\pi C_2^2. \tag{4.10}$$

Our simulations below confirm the accuracy of this prediction.

It is natural to undo the nondimensionalization of the equations to seek a prediction for the period of oscillations of landslide motion in the field. However, the enormous ranges of D, K, and v_{ref} in Table 2.1 diminish the value of this exercise. Specifically, one obtains oscillation periods ranging from about 10^{-7} to 10^3 seconds. At the small end, these periods will be unobservable by conventional measurement techniques. At the upper end, these periods are similar to those sometimes observed in the field and also observed in the landslide experiments of Iverson et al. [6].

(b) *Numerical limitations.* Provided $\varepsilon C_1 \ll 1$, (4.9) characterizes the loss of stability in the PDE (2.23), (2.24) through Hopf bifurcation. However, this relation is not accurate for numerical simulations if

$$\varepsilon C_1 \leq \mathcal{O}(h^3), \tag{4.11}$$

where h is the mesh size. To motivate this assertion, first recall that the eigenvalues of the PDE at the bifurcation point have absolute value

$$\mu^2/2 = 1/(2C_2^2) = 2^{-1/3}(\varepsilon C_1)^{-2/3} \gg 1.$$

On the other hand, the largest eigenvalue of the discretization is $\mathcal{O}(h^{-2})$, and moreover the large eigenvalues of the discretization do not approximate eigenvalues of the PDE. These two observations warn of a mismatch if

$$(\varepsilon C_1)^{-2/3} \geq \mathcal{O}(h^{-2}),$$

which is equivalent to (4.11).

Let us illustrate this phenomenon for a second-order explicit discretization of the PDE. (In the simulations below, we used the Crank–Nicholson method, for which the analysis is similar in spirit but more technical in detail.) For a positive integer N, let $h = 1/N$ be the mesh size in discretizing space, and let

$$y_n = 1 - nh, \quad n = 0, 1, \ldots, N+1.$$

Consider the semidiscrete approximation for the pressure equation (2.23),

(4.12)

$$\begin{array}{llcll}
\text{(a)} & \partial_t p_n & = & h^{-2}\left[p_{n+1} - 2p_n + p_{n-1}\right], & n = 1, 2, \ldots, N, \\
\text{(b)} & p_0 & = & 0, & \\
\text{(c)} & (2h)^{-1}\left[p_{N-1} - p_{N+1}\right] & = & \psi v, &
\end{array}$$

and for the velocity equation (2.24),

$$\varepsilon \partial_t v = B_1 p_N + B_2 v. \tag{4.13}$$

As with the PDE, we look for solutions of (4.12), (4.13) such that $p_n(t)$ and $v(t)$ have exponential time dependence $e^{-\lambda t}$. It follows from (4.12)(a),(b) that

$$p_n(t) = Pe^{-\lambda t}(z^n - z^{-n}), \qquad n = 0, 1, \ldots, N+1,$$

$$v(t) = Ve^{-\lambda t},$$

where

$$\lambda = \frac{z - 2 - z^{-1}}{h^2}. \tag{4.14}$$

Substituting into (4.12)(c), (4.13), we obtain a solution provided

$$\det \begin{bmatrix} (2h)^{-1}(z - z^{-1}) & \psi \\ B_1 & B_2 - \varepsilon h^{-2}(z - 2 - z^{-1}), \end{bmatrix} = 0, \tag{4.15}$$

where we have divided the first column of this determinant by z^N.

We analyze the Hopf bifurcation in this system as with the PDE: i.e., we ask when there is a solution z of (4.15) such that λ, computed according to (4.14), lies on the imaginary axis, say, $\lambda = i\mu$. We consider only the asymptotic range $\mu \gg 1$. Hence $z \gg 1$, and we may solve (4.14) for z approximately by neglecting z^{-1}; i.e.,

$$z = 2 + i\mu h^2.$$

Substituting this approximation into (4.15) and solving for B_i from (3.3), we rewrite (4.15) as

$$\det \begin{bmatrix} h^{-1} + i\mu h/2 & \psi \\ 1/(\psi C_1) & C_2/C_1 - i\varepsilon\mu \end{bmatrix} = 0. \tag{4.16}$$

From the vanishing of the imaginary part of this equation, we conclude that

$$C_2 = \frac{2}{h^2}\, \varepsilon C_1, \tag{4.17}$$

which is the relation that characterizes Hopf bifurcation at large eigenvalues in the discretization; thus, (4.17) replaces (4.8) when discretization effects invalidate the latter.

5. Supporting computations.

5.1. Methods. To test our bifurcation predictions and examine details of the dynamics described by (2.2) and (2.7), we solved discretized versions of the equations numerically. Values of most parameters used in the computations were fixed to match those of physical landslide experiments in which both quasi-steady sliding and stick-slip behavior were observed [6], whereas values of K, D, and v_{ref} were modified systematically to make computational transects of the εC_1-C_2 parameter space (Table 2.1). Specifically, values of ε were manipulated by adjusting the values of K and D in accordance with values appropriate for diverse soils; then, while holding εC_1 essentially constant, C_2 was increased incrementally through a plausible range by adjusting v_{ref}. As C_2 increased, bifurcation was detected as a transition from convergent oscillations (leading to a stable steady state) to divergent oscillations (leading to repetitive stick-slip cycles) in the v-p phase plane, a result described in more detail below.

With few exceptions, the value of the rate-weakening friction parameter a used in all computations was 0.02, consistent with observations in rate-controlled shear tests with many soil-like materials [8, 12]. To attain values of C_2 large enough to cause bifurcation when $\varepsilon C_1 > 0.1$, however, it was necessary to increase a to 0.04. Such large values of a, C_2, and εC_1 are atypical and perhaps even physically implausible, but mathematically they characterize the upper fringes of the εC_1-C_2 parameter space.

Our computational algorithm employed an explicit fourth-order Runge–Kutta method to solve the ODE (2.7) and the Crank–Nicholson method to solve the PDE (2.2) [9] in an operator-splitting scheme. First a Runge–Kutta time step Δt was taken to advance the slide-block velocity $v_x(t)$ while holding the excess basal pore pressure $p_{\text{ex}}(0, t)$ constant; then using the new $v_x(t)$ to update the basal boundary condition (2.3b) a Crank–Nicholson time step Δt was taken to advance the pore-pressure diffusion solution $p_{\text{ex}}(y, t)$. The updated pore-pressure solution provided the basal pore pressure necessary to take the next Runge–Kutta step. Refinement of this scheme by using the mean $v_x(t)$ between successive time steps to update (2.3b) and then recompute $p_{\text{ex}}(y, t)$ using this mean yielded solutions that differed negligibly from those of the basic scheme, provided that time steps were sufficiently small. Therefore, we used the basic scheme for all computations reported in this paper.

Our discretization of (2.2) and (2.7) used times steps with a size Δt suitable for resolving slide-block acceleration, which had an intrinsic timescale K/g typically much smaller than that of pore-pressure diffusion (i.e., $\varepsilon \ll 1$). In trial calculations we initially set $\Delta t = (K/g) \sin \theta$, the time necessary for the block to accelerate from 0 to K in the absence of friction and pore-pressure feedback. Subsequent trials showed that when friction and feedback were present, $\Delta t \gg (K/g) \sin \theta$ could generally be used with negligible loss of accuracy. Therefore, we consistently employed $\Delta t = 0.0002s$ to produce all computational results reported in this paper, although we regularly checked these results against those obtained using smaller time steps. Also, for the sake of consistency, our spatial discretization of (2.2) employed $h = 0.001$ for all results reported here, except for trials exploring finite-h effects.

In all computations we used $H_w = 0.3701m$, a value 1% larger than the static limiting equilibrium value that applies when (2.9) reduces to an equality for the parameter values listed in Table 2.1. As indicated by (2.1), fixing the initial value of H_w also fixed the background pore-pressure distribution $p_{\text{hydro}}(y)$. An initial excess pore-pressure distribution could be specified by (2.18), using (2.20) as an estimate for the steady-state velocity. In practice, during production runs, we iterated (2.19) one or more times to improve the initial estimate (2.20), in order to hasten convergence to the steady state.

5.2. Results. Our computational results are summarized in Figure 5.1, a graph that depicts theoretical bifurcation curves and computed bifurcation points in the εC_1-C_2 parameter space. The broad range of values spanned by this parameter space reflects the broad range of K and D values that are physically plausible for diverse soils (e.g., [2]), and it illustrates the wide scope of the bifurcation phenomenon.

We determined all bifurcation points shown in Figure 5.1 to at least two significant digits. At this level of precision, the computed bifurcation points lie exactly on the theoretical Hopf-bifurcation curve Γ (solid line in Figure 5.1), provided the simulation is not polluted by finite-h effects. When $\varepsilon C_1 \leq \mathcal{O}(h^3)$, numerical effects determine the location of the bifurcation point, and it may be seen from the figure that the bifurcation is accurately described by (4.17) (dashed line in Figure 5.1).

The physical character of the Hopf bifurcation is illustrated by phase portraits depicting coevolution of $v_x(t)$ and $p_{\text{ex}}(0, t)$. Figure 5.2 shows typical phase portraits

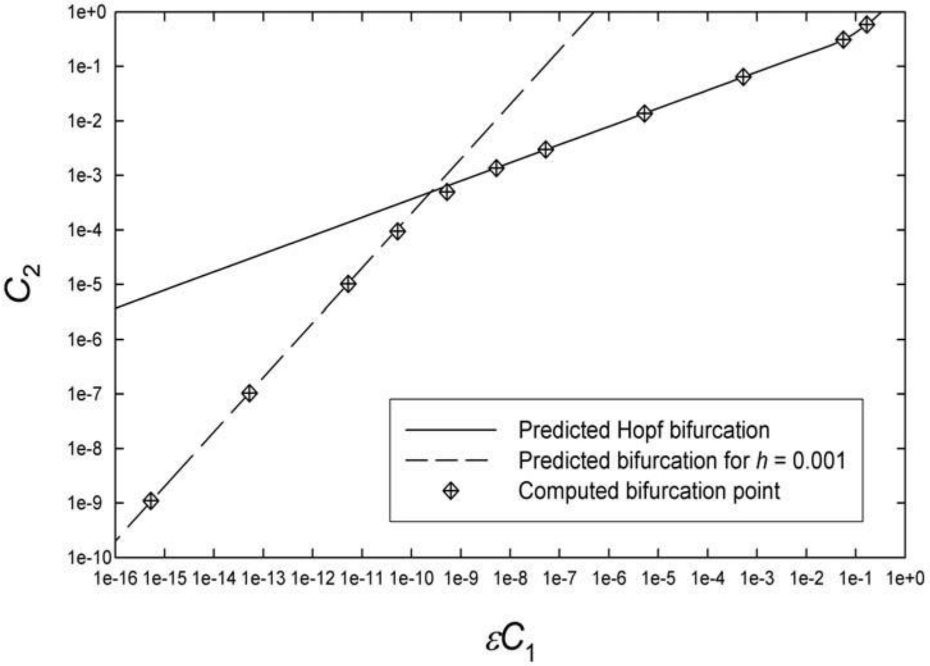


FIG. 5.1. *Comparison of computed bifurcation points with the Hopf bifurcation curve* Γ *predicted by* (4.6), (4.7) *(solid line), and the finite-h limit* (4.17) *for the case in which* $h = 0.001$ *(dashed line).*

for C_2 near the bifurcation point. The bifurcation is subcritical—when $C_2 > C_{2\text{Hopf}}$, the solution evolves to a periodic solution whose amplitude does not tend to zero as $C_2 \to C_{2\text{Hopf}}$. (Note that Figures 5.2B and 5.2C have different scales.) As expected, the evolution toward or away from the steady state slows down as the bifurcation point is approached [11].

For the parameter values used in making Figure 5.2, (4.10) predicts that near the bifurcation point, the period of oscillations is 0.325 s. By comparision, the oscillations depicted in Figure 5.2C have a computed period of 0.336 s. The fairly large discrepancy between these two numbers is related to the fact that the bifurcation is subcritical. Strictly speaking, (4.10) predicts the period of the small-amplitude, unstable orbits close to the bifurcation point, while Figure 5.2C shows a moderate-amplitude, stable orbit to which the solution jumps when C_2 exceeds $C_{2\text{Hopf}}$. A better comparison is provided by the nearly periodic, decaying solution shown in Figure 5.2B, in which the numerically estimated period is 0.324 s.

During part of the periodic orbits in Figures 5.2C and 5.2D, the velocity vanishes. This stick-slip behavior occurs because the nonlinearity limiting growth of the oscillations is a nonsmooth one, i.e., the discontinuous behavior of friction at $v = 0$. In no instance did oscillations persist without stick-slip behavior.

Transitions in phase-portrait behavior for other values of εC_1 were qualitatively similar to those illustrated in Figure 5.2 except for unphysically large values of εC_1: i.e., near the end of Γ at the point $(1/3, 1)$. In the latter case, the large-time orbit differs from Figure 5.2 in that the block "sticks" during a large fraction of the period, even immediately after bifurcation.

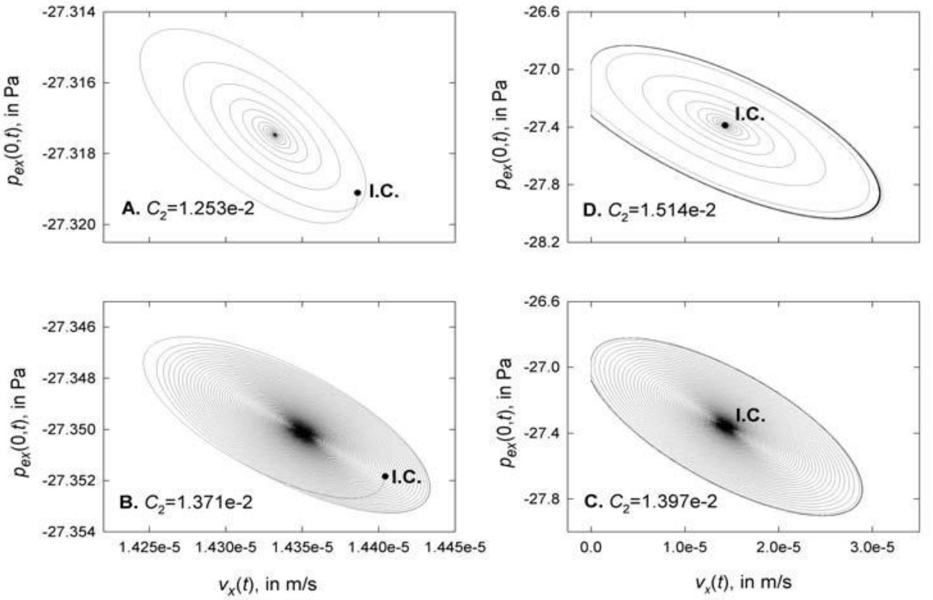


FIG. 5.2. *Examples of computed orbits in the* v_x, p_{ex} *phase plane for a case in which* $\varepsilon C_1 = 5.256\times10^{-6}$ *and* $C_{2Hopf} = 1.380\times10^{-2}$. *These computations used* $K = 2\times10^{-4}$ *m/s,* $D = 1\times10^{-3}$ *m*^{2}*/s,* $a = 0.02$, *and values of* v_{ref} *ranging from* 2.4×10^{-3} *to* 2.9×10^{-3} *m/s to obtain varying values of* C_2. *All other parameters were held fixed at the values listed in Table* 2.1. *The point marked "I.C." indicates the initial condition used in each computation. Note that the value of* C_2 *increases moving counterclockwise in the figure from frames* A *through* D. *In frame* B, C_2 *is just below critical, while in frame* C, *it is just above. Incidentally, in frame* D, *motion along the portion of the trajectory where* $v \equiv 0$, *which is governed by diffusion alone, is slower than along the remainder of the trajectory, and this effect is more pronounced if* C_2 *is further from the bifurcation point.*

6. Concluding discussion. Hopf bifurcation occurs in solutions of equations that provide a parsimonious model of landslide motion regulated by dilatancy, pore-pressure feedback, and rate-weakening friction. The bifurcation is manifested as an abrupt transition (thus, the bifurcation is subcritical) from stable, steady, downslope motion to periodic motion characterized by repetitive stick-slip cycles. The existence of stick-slip behavior in this system is noteworthy because, unlike classical models that exhibit stick-slip, our model includes no elastic element that exerts a variable and reversible driving force. (The archetype model for stick-slip behavior is a rate-weakening friction block pulled along a plane by an elastic spring.) Instead, in our model, the driving force is the steady pull of gravity, and the frictional resisting force is mediated by pore-pressure diffusion. Effects of pore-pressure diffusion have also been studied in the context of stick-slip models that include an elastic driving element [10], but to our knowledge no previous model has duplicated ours in omitting elastic forces while retaining the capacity for stick-slip behavior.

Analysis and computations show that the Hopf bifurcation leading to stick-slip behavior in our model is precisely governed by the parameters εC_1 and C_2, although decomposition of these parameters into their physical components shows that their variation depends mostly on variations in ε and the velocity ratio K/v_{ref}. Physically, the timescale ratio ε specifies the relative speeds at which the landslide characteristically moves and excess pore pressure characteristically diffuses, whereas K/v_{ref}

specifies the degree to which rate-weakening friction affects landslide motion. For relevant parameter values (Table 2.1), $\varepsilon \ll 1$ applies almost universally, indicating that pore-pressure diffusion is a relatively slow process that serves to regulate the inherently faster process of landslide motion. Also, for relevant parameter values, $K/v_{\text{ref}} \ll 1$ is typical, although K/v_{ref} values of order 1 or larger are possible (see Table 2.1). For $K/v_{\text{ref}} \ll 1$, the characteristic slip velocity $v_x = \mathcal{O}(K)$ is small. Taking the limit $v_{\text{ss}} = 0$ and then combining (3.3), (2.25), and (2.6), we calculate that

$$C_2 = \frac{\sin\theta}{\psi A_2}\,\frac{aK}{2\mu_0 v_{\text{ref}}},$$

which shows that, for $K/v_{\text{ref}} \ll 1$, increases in K/v_{ref} produce increases in C_2 and hence decreases in stability. This is the behavior typically observed in our computations. By contrast, if K/v_{ref} becomes comparable to or greater than 1, increases in K/v_{ref} can produce decreases in C_2 and hence *increases* in stability. Physically, this behavior reflects the fact that friction, which decreases logarithmically at large velocities, becomes increasingly insensitive to v as slip rates grow large.

Another important observation regarding the physics described by our model concerns the consistent manner in which orbits in the v_x-p_{ex} phase plane are skewed. As shown in Figure 5.2, the minimum $p_{\text{ex}}(0,t)$ always lags the maximum $v_x(t)$ by less than one quarter of an orbit cycle, irrespective of whether orbits diverge unstably or converge to a fixed point. Similar orbit skewness is exhibited in all of our computational results. However, as $\varepsilon C_1 \to 0$ the orbit skewness gradually diminishes, so that orbits become almost symmetrical about the line $v_x = v_{\text{ss}}$ and the phase lag approaches 1/4-cycle. This skewness of the orbits is a consequence of inertia. To illustrate this, observe that $\varepsilon C_1 \to 0$ if, for example, $K \to 0$; it follows from (2.10) that the dimensional steady-state velocity tends to zero as $K \to 0$, and hence inertial effects will disappear in this limit.

Finally, we emphasize that two key effects are not included in our model: (i) parameter evolution (e.g., dilatancy evolution) and (ii) a rate-and-state friction law in which the friction coefficient evolves with time [1, 10]. Such effects could lead to other kinds of instabilities, including a possibly more complex bifurcation than what we have analyzed.

Appendix.

PROPOSITION A.1. *If $C > 0$, the function*

$$f(\lambda) \equiv \frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} - C\lambda$$

has no zeros in the closed left half plane $\{\lambda : \operatorname{Re}\lambda \le 0\}$.

Proof. For any $R > 0$, let Ω_R be the half disk

$$\{\lambda : \operatorname{Re}\lambda < 0,\ |\lambda| < R\},$$

inside which f is analytic. According the principle of the argument [7, Chapter 4, section 4], provided f is nonzero on $\partial\Omega_R$, the number of zeros of f in Ω_R equals the variation of $\arg f$ around the boundary. It is obvious that $\operatorname{Re}(-C\lambda) \ge 0$ on $\partial\Omega_R$. We claim that, no matter how large R may be,

$$\operatorname{Re}\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} > 0 \qquad \text{on } \partial\Omega_R.$$

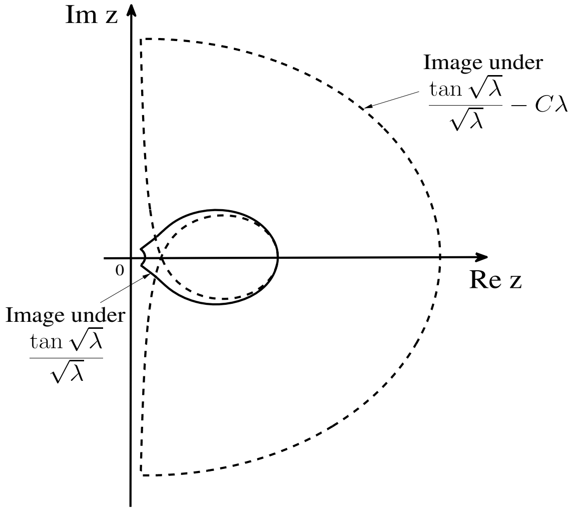


FIG. A.1. *Images of the semicircle $\partial\Omega_R$ in the complex plane under two mappings (see the appendix). The radius $R = 100$, and the constant $C = 0.02$.*

Thus, the variation of the argument of f around $\partial\Omega_R$ is zero, and this proves the result.[3]

Let us prove the claim. For λ on the semicircle $\{\lambda : \operatorname{Re}\lambda \le 0, |\lambda| = R\}$, we invoke the asymptotic form (4.3) and observe that

$$\operatorname{Re}\frac{i}{\sqrt{\lambda}} = |\lambda|^{-1/2}\cos\left(\frac{\pi - \arg\lambda}{2}\right) > 0;$$

the reason for the inequality is that $-\pi/4 \le (\pi - \arg\lambda)/2 \le \pi/4$. For λ on the imaginary axis, since $f(\overline{\lambda}) = \overline{f(\lambda)}$, it suffices to restrict our attention to $\{\operatorname{Im}\lambda \ge 0\}$. Along the nonnegative imaginary axis we may parametrize $\sqrt{\lambda}$ as

$$\sqrt{\lambda} = (1+i)t, \qquad t \ge 0.$$

Substituting into (4.4), we find

$$\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} = \frac{i}{(1+i)t}\,\frac{1 - e^{2i(1+i)t}}{1 + e^{2i(1+i)t}}.$$

Multiplying both the numerator and the denominator by the complex conjugate of the denominator and collecting terms, we calculate that

$$\operatorname{Re}\frac{\tan\sqrt{\lambda}}{\sqrt{\lambda}} = p(t)\,\frac{\sinh 2t + \sin 2t}{t} > 0,$$

where the positive factor

$$p(t) = \left|1 + e^{2i(1+i)t}\right|^{-2}$$

comes from the modulus squared of the denominator. This proves the proposition. □

[3] The reader may find it interesting to consult Figure A.1, which shows the image of $\partial\Omega_R$ under $(\tan\sqrt{\lambda})/\sqrt{\lambda}$ and under f.

REFERENCES

[1] J. H. Dieterich and B. D. Kilgore, *Direct observation of frictional contacts: New insights for state-dependent properties*, Pure and Applied Geophysics, 143 (1994), pp. 283–302.

[2] R. A. Freeze and J. A. Cherry, *Groundwater*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[3] M. Golubitsky and D. Schaeffer, *Singularities and Groups in Bifurcation Theory*, *Vol.* I, Springer-Verlag, New York, 1985.

[4] R. M. Iverson, *Dilatancy, pore-pressure feedback, and regulation of landslide (and thrust fault) motion*, Geological Society of America Annual Meeting, Abstracts with Programs, 34 (2003), p. 196.

[5] R. M. Iverson, *Regulation of landslide motion by dilatancy and pore pressure feedback*, J. Geophys. Res., 110 (2005), F02015.

[6] R. M. Iverson, M. E. Reid, N. R. Iverson, R. G. LaHusen, M. Logan, J. E. Mann, and D. L. Brien, *Acute sensitivity of landslide rates to initial soil porosity*, Science, 290 (2000), pp. 513–516.

[7] N. Levinson and R. Redheffer, *Complex Variables*, Holden-Day, San Francisco, CA, 1970.

[8] P. L. Moore, N. R. Iverson, and R. M. Iverson, *Frictional properties of the Mount St. Helens gouge*, Chapter 20 in A Volcano Rekindled: The Renewed Eruption at Mount St. Helens, 2004-2006, D. R. Sherrod, W. E. Scott, and P.E. Stauffer, eds., U.S. Geological Survey Professional Paper 1750, in press.

[9] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes*, Cambridge University Press, Cambridge, UK, 1986.

[10] P. Segall and J. R. Rice, *Dilatancy, compaction, and slip instability of a fluid-infiltrated fault*, J. Geophys. Res., 100 (1995), pp. 22155–22172.

[11] S. Strogatz, *Nonlinear Dynamics and Chaos*, Westview Press, Boulder, CO, 1994.

[12] T. E. Tika, P. R. Vaughan, and L. J. Lemos, *Fast shearing of preexisting shear zones in soil*, Geotechnique, 46 (1996), pp. 197–233.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 787–809

A CONTINUUM THEORY OF CHIRAL SMECTIC C LIQUID CRYSTALS[*]

MARIA-CARME CALDERER[†] AND SOOKYUNG JOO[‡]

Abstract. We formulate a nonlinear continuum theory of flow of chiral smectic C liquid crystals (C*) involving molecular director, layer order parameter, polarization vector, flow velocity, and hydrostatic pressure fields. In addition to chiral orientational ordering, smectic C* phases also present positional ordering, with molecular centers of mass arranged in one dimensional layers. The nonzero tilt angle of the molecular director with respect to the layer normal together with the chirality is responsible for the ferroelectric nature of the phase. This results in a stronger coupling with applied electric fields than the dielectric nematic. We apply the model to study the molecular reorientation dynamics in homeotropic geometry under the influence of an applied electric field. The switching process between states with opposite polarization is understood by the traveling wave solution of the system. We prove existence and uniqueness of the traveling wave and show that the predicted switching time is smaller than that when the flow effect is neglected. We also obtain bounds on the speed of switching and an optimality condition on the parameters of the problem. Numerical simulations confirm the predictions of the analysis.

Key words. continuum theory, smectic liquid crystals, molecular reorientation dynamics, ferroelectric liquid crystals, traveling wave

AMS subject classifications. 35Q35, 35Q51, 76D07, 80A17

DOI. 10.1137/070696477

1. Introduction.

We develop a model of smectic C* liquid crystals accounting for elastic, hydrodynamic, and electrostatic effects. The free energy includes the Oseen–Frank energy of nematic liquid crystals, the smectic C energy of the form proposed by Chen and Lubensky, and the ferroelectric electrostatic energy. We apply the governing equations to study the switching dynamics, in homeotropic geometry, between two states with opposite electric polarization. We apply a variational method to characterize the speed of the switching traveling wave and show that the predicted speed is greater than in the approach that neglects flow. We also obtain an optimality condition of the speed in terms of the parameters of the problem. We perform numerical simulations to illustrate the dynamics of switching. A main feature of our work is the study of the backflow effect due to the spontaneous polarization of the liquid crystal.

Liquid crystal phases form when a material has a degree of positional or orientational ordering yet stays in a liquid state. In the nematic state, molecules tend to align themselves along a preferred direction with no positional order of centers of mass. The unit vector field $\mathbf{n}$, nematic director, represents the average direction of molecular alignment. Moreover, if the liquid crystal is chiral, $\mathbf{n}$ follows a helical pattern with temperature-dependent pitch. Upon lowering the temperature, or increasing concentration, according to whether the liquid crystal is thermotropic or lyotropic, the nematic liquid crystal experiences a transition to the smectic A phase

[*] Received by the editors July 4, 2007; accepted for publication (in revised form) September 3, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/siap/69-3/69647.html

[†] School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (mcc@math.umn.edu). This author's research was supported by NSF-FRG grant DMS-0456232.

[‡] Mathematics Department, University of California, Santa Barbara, CA 93106 (sjoo@math.ucsb.edu).

with molecules arranged along equally spaced layers. The molecules tend to align themselves along the direction perpendicular to the layers. Upon transition to the lower temperature smectic C phase, a symmetry break occurs, with molecules making a nonzero tilt angle with the layer normal. Values of the tilt angle α are found to be between 20 and 35 degrees and depend on the material and temperature.

We consider chiral smectic C liquid crystals and label them C*, according to conventional notation. One relevant feature of liquid crystal molecules that form smectic C* phases is the presence of a *side chain* giving a transverse electric dipole and therefore yielding the polarization field $\mathbf{P}$ of the theory. Level surfaces of the scalar variable ϕ describe smectic layers. A schematic representation of the fields of a smectic C* phase is given in Figure 1, where the normal vector to the layers corresponds to the axis of a cone of semiangle α, with $\mathbf{n}$ being allowed to rotate on the surface. The polarization vector $\mathbf{P}$ is perpendicular to both the layer normal $\nabla\phi$ and the director field [16]. Additional fields of the hydrodynamic theory are the velocity field $\mathbf{u}$ and the hydrostatic pressure p. The latter is the constraint associated with the assumption of fluid incompressibility.

In chiral configurations, since $\mathbf{P}$ rotates with $\mathbf{n}$, the net polarization of the material is zero. Therefore, ferroelectric states correspond to configurations with constant $\mathbf{n}$. The electrostatic effects due to polarization dominate the dielectric effects of standard nematic liquid crystals. Consequently, faster switching devices are achieved with smectic C* liquid crystals.

In this article, we study the transition between states with opposite polarization and determine lower bounds for the speed of the connecting traveling wave. The latter corresponds to a chiral configuration with periodically varying $\mathbf{n}$ and $\mathbf{P}$. The switching takes place upon reversing the direction of the applied electric field. The stability of the polarized states was studied in [23]. The traveling wave of our problem represents the backflow effect, that is, the flow generated by changes in the applied electric field.

The free energy density of the model consists of nematic, smectic C, and electrostatic contributions. The form of the smectic C free energy, $\mathbf{F}_S$, that we study was introduced by Chen and Lubensky in 1976, based on the Landau–de Gennes model for smectic A [6]. They investigated the nematic to smectic phase transition, and it was later used by Renn and Lubensky to predict the twist grain boundary phase in cholesteric smectic [19]. However, the free energy density $\mathbf{F}_S$ is degenerate in that it lacks second order coercivity in the direction $\mathbf{n}$. In order to avoid the anisotropic quartic order derivatives in the Chen–Lubensky model, Luk'yanchuk proposed a modified model [20]. The new model was later used in [13] to rigorously analyze the temperature phase transition from chiral nematic to chiral smectic liquid crystals. The analysis of the ferroelectric smectic C* phases was carried out in [24], where the energy minimizers are further required to satisfy the electrostatic Maxwell equations. The hydrodynamic theory that we propose combines the approaches by Leslie and Ericksen (for details, see [8], [4], and [29]) for nematic and the work by W. E [9] for smectic A liquid crystals. The latter follows the model by Kleman and Parodi [14] also for smectic A phases, where the concepts of permeation force and molecular field were introduced as forces driving smectic A flow. However, since the layer position completely specifies the director field in smectic A, W. E shows that only the permeation force is responsible for the dynamics of smectic A liquid crystals. This is not the case for the smectic C modeling, where both forces are needed to describe the hydrodynamics. We use a variational approach together with the dissipation inequality to determine the elastic and viscous components, respectively, of such forces. Furthermore, the Lorenz force associated with the charge density $-\mathrm{div}\mathbf{P}$ enters the equation of balance of linear momentum.

Leslie, Stewart, and Nakagawa also developed a nonlinear continuum theory for smectic C liquid crystals, using the **c** director, which is the projection of **n** onto the layer, and the unit vector normal to the layers (see [18] and [29]). Their theory is constrained to exclude variations in the layer spacing thickness and changes in tilt with respect to the layer. The nonlinear continuum theory in the present paper is also restricted to the constant tilt angle case, excluding the variation of the tilt angle between the director and the layer normal as in [18]. However, our model allows the variation of the layer spacing thickness.

The second part of the paper is devoted to the study of the switching dynamics of a smectic C* sample confined between parallel bounding planes. We assume that the electric field is applied parallel to the smectic layers. We derive the governing equation of the director and the flow equation in the homeotropic geometry, where the smectic layers are parallel to the bounding plates. When the flow is neglected, the director profile can be understood by the traveling wave solution of the resulting nonlinear reaction diffusion equation (see [7], [21], [25], and [26]). This equation also represents the gradient flow of the energy. One main goal of our work is to study the traveling wave solutions, taking flow and ferroelectric effects into account and estimating the speed of the corresponding traveling wave. The variational characterization of the speed follows the approach in [2] for reaction-diffusion equations. Furthermore, we obtain an optimal lower bound of the speed in terms of the viscous and smectic parameters of the model. Numerical simulations of the problem explore ranges of parameters, from the case in which flow is neglected to cases in which parameters approach the optimal lower bound of the speed. We find a very good agreement with the predictions of the analysis.

Section 2 is devoted to static theory, and dissipation and hydrodynamics are discussed in section 3. The analysis of traveling waves of the switching problem and the corresponding numerical simulations are developed in section 4.

2. Hydrostatic theory.

2.1. Smectic C* free energy. The total free energy density consists of the nematic f_n and smectic f_s parts. The Oseen–Frank energy density for a nematic is given by

$$f_n = \frac{K_1}{2}(\nabla\cdot\mathbf{n})^2 + \frac{K_2}{2}(\mathbf{n}\cdot\nabla\times\mathbf{n}+\tau)^2 + \frac{K_3}{2}|\mathbf{n}\times(\nabla\times\mathbf{n})|^2,$$

where K_1, K_2, and K_3 are the splay, twist, and bend elastic constants, respectively. The parameter τ denotes the cholesteric twist.

In order to associate smectic and nematic structure with a state $(\mathbf{n},\Psi)$, we write

$$\Psi(\mathbf{x}) = \rho(\mathbf{x})e^{i\varphi(\mathbf{x})}.$$

Then the molecular mass density is defined by

$$\delta(\mathbf{x}) = \rho_0(\mathbf{x}) + \frac{1}{2}(\Psi(\mathbf{x})+\Psi^*(\mathbf{x})) = \rho_0(\mathbf{x}) + \rho(\mathbf{x})\cos\varphi(\mathbf{x}),$$

where ρ_0 is a locally uniform mass density, $\rho(\mathbf{x})$ is the mass density of the smectic layers, and φ parametrizes the layers so that $\nabla\varphi$ is the direction of the layer normal.

Now the smectic C energy density is given by

$$f_s = \frac{D}{2}|\mathbf{D}_\mathbf{n}^2\Psi|^2 - \frac{C_\perp}{2}|\mathbf{D}_\mathbf{n}\Psi|^2 + \frac{C_\|}{2}|\mathbf{n}\cdot\mathbf{D}_\mathbf{n}\Psi|^2,$$

where $\mathbf{D}_\mathbf{n} = \nabla - iq\mathbf{n}$, $\mathbf{D}_\mathbf{n}^2 = \mathbf{D}_\mathbf{n}\cdot\mathbf{D}_\mathbf{n}$, and $D, C_\perp, C_\|$ are positive constants. The model for smectic C energy was proposed by Chen and Lubensky [6], but we use the modified model, introduced by Luk'yanchuk [20]. Since we investigate smectic structure far from the nematic–smectic transition, we assume that the magnitude of the smectic order parameter is a constant. We may assume that $\Psi = e^{i\varphi}$. Then f_s becomes

$$\begin{aligned} f_s &= \frac{D}{2}|\nabla\varphi - q\mathbf{n}|^4 + \frac{D}{2}(\Delta\varphi - q\nabla\cdot\mathbf{n})^2 - \frac{C_\perp}{2}|\nabla\varphi - q\mathbf{n}|^2 + \frac{C_\|}{2}(\mathbf{n}\cdot\nabla\varphi - q)^2 \\ &= \frac{D}{2}\left(|\nabla\varphi - q\mathbf{n}|^2 - \frac{C_\perp}{2D}\right)^2 + \frac{C_\|}{2}(\mathbf{n}\cdot\nabla\varphi - q)^2 + \frac{D}{2}(\Delta\varphi - q\nabla\cdot\mathbf{n})^2. \end{aligned}$$

If $\mathbf{n}$ is a constant and φ is linear, then we can see that the energy is minimized if and only if $\mathbf{n}\cdot\nabla\varphi = q$ and $|\nabla\varphi - q\mathbf{n}|^2 = \frac{C_\perp}{2D}$. This corresponds to a uniform smectic C state with tilt angle α, between the director and the layer normal, determined by $\tan^2\alpha = C_\perp/(2Dq^2)$ and layer thickness d satisfying $(\frac{2\pi}{d})^2 = q^2 + \frac{C_\perp}{2D}$.

We get the free energy density

$$f_d = f_n + f_s.$$

Note that there are two constraints:

$$|\mathbf{n}| = 1 \qquad \text{and} \qquad \mathbf{n}\cdot\nabla\varphi = \cos\alpha|\nabla\varphi|. \tag{2.1}$$

We consider the total smectic C free energy density

$$\tilde{f} = f_d + f_l, \tag{2.2}$$

where the last term is present in order to make use of Lagrange multipliers:

$$f_l = \frac{\lambda}{2}(\mathbf{n}\cdot\mathbf{n} - 1) + \beta(\mathbf{n}\cdot\nabla\varphi - \cos\alpha|\nabla\varphi|).$$

2.2. The molecular field and the permeation force. In the nematic, the molecular field can be obtained through the deformation of the director field while the centers of gravity of the molecules are fixed. On the other hand, the directors are parallel to the layer normal in smectic A. As a result, W. E instead discussed the permeation forces, the normal forces acting on layers in [9]. In smectic C, the directors are tilted with respect to the layer normal, and hence both the molecular field and the permeation force need to be discussed. First we obtain the equilibrium conditions in bulk by writing the variation of the total free energy with respect to the director and the layer normal variations, while keeping the material undeformed. Let D be any region inside the liquid crystal. We get

$$\delta\int_D \tilde{f} = \int_D \left(\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\delta(\Delta\varphi) + \frac{\partial\tilde{f}}{\partial(\partial_i\varphi)}\delta(\partial_i\varphi) + \frac{\partial\tilde{f}}{\partial(\partial_i\mathbf{n}_j)}\delta(\partial_i\mathbf{n}_j) + \frac{\partial\tilde{f}}{\partial\mathbf{n}_i}\delta\mathbf{n}_i\right) d\mathbf{x}$$

$$
\begin{aligned}
&= \int_D \left(\partial_i \left[\partial_i \left(\frac{\partial \tilde{f}}{\partial(\Delta\varphi)} \right) - \frac{\partial \tilde{f}}{\partial(\partial_i \varphi)} \right] \delta\varphi + \left[\frac{\partial \tilde{f}}{\partial \mathbf{n}_i} - \partial_j \left(\frac{\partial \tilde{f}}{\partial(\partial_j \mathbf{n}_i)} \right) \right] \delta\mathbf{n}_i \right) d\mathbf{x} \\
&+ \int_{\partial D} \left(\frac{\partial \tilde{f}}{\partial(\Delta\varphi)} \partial_j(\delta\varphi) + \left[\frac{\partial \tilde{f}}{\partial(\partial_j \varphi)} - \partial_j (\frac{\partial \tilde{f}}{\partial(\Delta\varphi)}) \right] \delta\varphi + \frac{\partial \tilde{f}}{\partial(\partial_j \mathbf{n}_i)} \delta\mathbf{n}_i \right) \nu_j \, ds \\
&=: \int_D (-g\delta\varphi - \mathbf{h}_i \delta\mathbf{n}_i) \, d\mathbf{x} + \int_{\partial D} \left(\frac{\partial \tilde{f}}{\partial(\Delta\varphi)} \partial_j(\delta\varphi) - \tau_j \delta\varphi + \pi_{ij} \delta\mathbf{n}_i \right) \nu_j \, ds,
\end{aligned}
$$

where

$$
\begin{aligned}
g &= -\nabla \cdot \tau \\
&= -\nabla \cdot \left[\nabla \left(\frac{\partial f_d}{\partial(\Delta\varphi)} \right) - \frac{\partial f_d}{\partial(\nabla\varphi)} - \beta \left(\mathbf{n} - \cos\alpha \frac{\nabla\varphi}{|\nabla\varphi|} \right) \right], \\
\mathbf{h}_i &= -\frac{\partial f_d}{\partial \mathbf{n}_i} + \partial_j \pi_{ij} - \lambda \mathbf{n}_i - \beta \partial_i \varphi,
\end{aligned}
\tag{2.3}
$$

using the notation

$$
\pi_{ij} = \frac{\partial f_d}{\partial(\partial_j \mathbf{n}_i)}. \tag{2.4}
$$

2.3. The elastic stress. We now calculate the elastic stress associated with the infinitesimal deformation of the body, while holding the location of the layers and the director field fixed. For this, we let

$$
\begin{aligned}
&\mathbf{r}' = \mathbf{r} + \mathbf{u}(\mathbf{r}), \\
&\mathbf{n}'(\mathbf{r}') = \mathbf{n}'(\mathbf{r} + \mathbf{u}) = \mathbf{n}(\mathbf{r}), \\
&\varphi'(r') = \varphi(\mathbf{r} + \mathbf{u}) = \varphi(\mathbf{r}).
\end{aligned}
$$

Using the relations

$$
\frac{\partial r_i'}{\partial r_j} = \delta_{ij} + \frac{\partial u_i}{\partial r_j} \quad \text{and} \quad \frac{\partial r_i}{\partial r_j'} \simeq \delta_{ij} - \frac{\partial(\delta u_i)}{\partial r_j},
$$

we get

$$
\frac{\partial \varphi'}{\partial r_j'} \simeq \frac{\partial \varphi}{\partial r_j} - \frac{\partial \varphi}{\partial r_k} \frac{\partial u_k}{\partial r_j} \quad \text{and} \quad \frac{\partial \mathbf{n}_i'}{\partial r_j'} \simeq \frac{\partial \mathbf{n}_i}{\partial r_j} - \frac{\partial \mathbf{n}_i}{\partial r_k} \frac{\partial u_k}{\partial r_j}.
$$

Hence

$$
\begin{aligned}
\delta(\partial_i \varphi) &\simeq -\frac{\partial \varphi}{\partial r_k} \frac{\partial u_k}{\partial r_i}, \\
\delta(\partial_i \mathbf{n}_j) &\simeq -\frac{\partial \mathbf{n}_j}{\partial r_k} \frac{\partial u_k}{\partial r_i}, \\
\delta(\Delta\varphi) &\simeq -2(\partial_{ik}\varphi)(\partial_i u_k) - (\partial_k \varphi)\Delta u_k.
\end{aligned}
$$

Taking these approximations into account, we now calculate the corresponding variation of the energy of a subdomain D in Ω. For this, we use integration by parts. This gives

$$
\begin{aligned}
\delta\int_D \tilde{f} &= \int_D \left(\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\delta(\Delta\varphi) + \frac{\partial\tilde{f}}{\partial(\partial_i\varphi)}\delta(\partial_i\varphi) + \frac{\partial\tilde{f}}{\partial(\partial_j\mathbf{n}_i)}\delta(\partial_j\mathbf{n}_i)\right) d\mathbf{x}\\
&= \int_D \left(\frac{\partial f_d}{\partial(\Delta\varphi)}\left[-2(\partial_{jk}\varphi)(\partial_j\mathbf{u}_k) - (\partial_k\varphi)(\partial_j^2\mathbf{u}_k)\right]\right) d\mathbf{x}\\
&+ \int_D \left(\left[\frac{\partial f_d}{\partial(\partial_j\varphi)} + \beta\left(\mathbf{n}_j - \cos\alpha\frac{\partial_j\varphi}{|\nabla\varphi|}\right)\right]\left(-(\partial_k\varphi)(\partial_j\mathbf{u}_k)\right)\right) d\mathbf{x}\\
&+ \int_D \left(\frac{\partial f_d}{\partial(\partial_i\mathbf{n}_j)}\left(-(\partial_k\mathbf{n}_i)(\partial_j\mathbf{u}_k)\right)\right) d\mathbf{x} - \int_{\partial D}\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}(\partial_k\varphi)(\partial_j\mathbf{u}_k)\nu_j\, ds\\
&=: \int_D \left(\sigma^d_{kj}(\partial_j\mathbf{u}_k)\right) d\mathbf{x} - \int_{\partial D}\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}(\partial_k\varphi)(\partial_j\mathbf{u}_k)\nu_j\, ds,
\end{aligned}
$$

where

$$
\sigma^d_{kj} = \left[-\frac{\partial f_d}{\partial(\Delta\varphi)}\partial_{jk}\varphi + \tau_j(\partial_k\varphi) - \pi_{ij}(\partial_k\mathbf{n}_i)\right] \tag{2.5}
$$

is the deviatoric part of the stress tensor. To take this incompressibility constraint into account, we modify the previous calculations to include the corresponding Lagrange multiplier term. For this, let us consider the free energy density

$$
f = \tilde{f} - p\nabla\cdot\mathbf{u},
$$

where p is a Lagrange multiplier. This leads to a modified elastic stress

$$
\sigma^e_{kj} = \sigma^d_{kj} - p\delta_{kj}. \tag{2.6}
$$

2.4. The equilibrium equations. By combining all variations from the previous sections, we have the total variation of $\tilde{f}$:

$$
\begin{aligned}
\delta\int_D \tilde{f} = \int_D &\left(\sigma^e_{kj}(\partial_j\mathbf{u}_k) - \mathbf{h}_k\delta\mathbf{n}_k - g\delta\varphi\right) d\mathbf{x}\\
+ \int_{\partial D} &\left(\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\partial_j(\delta\varphi) - \tau_j\delta\varphi + \pi_{kj}\delta\mathbf{n}_k - \frac{\partial\tilde{f}}{\partial(\Delta\varphi)}(\partial_k\varphi)(\partial_j\mathbf{u}_k)\right)\nu_j\, ds.
\end{aligned} \tag{2.7}
$$

By integration by parts, it becomes

$$
\begin{aligned}
\delta\int_D \tilde{f} = \int_D &\left(-\partial_j(\sigma^e_{kj})\mathbf{u}_k - \mathbf{h}_k\delta\mathbf{n}_k - g\delta\varphi\right) d\mathbf{x}\\
+ \int_{\partial D} &\left(\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\partial_j(\delta\varphi) - \tau_j\delta\varphi + \pi_{kj}\delta\mathbf{n}_k\right)\nu_j\, ds\\
+ \int_{\partial D} &\left(\sigma^e_{kj}\mathbf{u}_k - \frac{\partial\tilde{f}}{\partial(\Delta\varphi)}(\partial_k\varphi)(\partial_j\mathbf{u}_k)\right)\nu_j\, ds.
\end{aligned}
$$

Then the hydrostatic equilibrium condition is

$$
\delta\int_D \tilde{f} = 0
$$

for all $D \subseteq \Omega$ and for admissible variations. Taking variations such that $\delta \mathbf{n} = 0$, $\delta\varphi = 0$, and $\mathbf{u}$ and its gradient vanish at the boundary gives the system of partial differential equations

$$\mathbf{h} = 0, \qquad g = 0, \tag{2.8}$$

and

$$\partial_j \sigma_{kj} = 0. \tag{2.9}$$

These, together with two constraint relations (2.1), give the equilibrium equations for smectic C without the external fields. In three dimensions, the system (2.1) and (2.8) consists of six scalar equations for the six unknowns $\mathbf{n}$, φ, λ, and β. Well posedness of this system, in its variational form, was studied in [13] and [24]. Moreover, in [13], the authors performed an extensive phase transition and stability analysis of equilibrium states. In [24], the role of permanent polarization was a main focus of the work.

Notice that the combined system (2.1), (2.8), and (2.9) is overdetermined. This issue in nematic liquid crystals was addressed by Ericksen in [10]. He argues that artificial body forces have to be included in the equations for the system to be closed.

2.5. The balance of torques. We integrate by parts (2.7) to obtain

$$\begin{aligned} \delta \int_D \tilde{f} &= \int_D \left(\sigma^e_{kj}(\partial_j \mathbf{u}_k) - \mathbf{h}_k \delta \mathbf{n}_k - \tau_k \partial_k(\delta\varphi)\right) d\mathbf{r} \\ &\quad + \int_{\partial D} \left(\frac{\partial \tilde{f}}{\partial(\Delta\varphi)} \partial_j(\delta\varphi) + \pi_{kj}\delta\mathbf{n}_k - \frac{\partial \tilde{f}}{\partial(\Delta\varphi)}(\partial_k\varphi)(\partial_j \mathbf{u}_k)\right) \nu_j \, ds. \end{aligned} \tag{2.10}$$

Notice that $\tilde{f}$ is invariant under the rotation of the centers of gravity, the directors, and the layers by the same angle ω. Now the energy is unchanged under the following replacement:

$$\begin{aligned} \mathbf{u}(\mathbf{r}) &= \omega \times \mathbf{r}, \\ \delta\mathbf{n}(\mathbf{r}) &= \omega \times \mathbf{n}, \\ \delta\varphi &= 0, \end{aligned}$$

where ω is the rotation vector. Then we have

$$\partial_j \mathbf{u}_k = \varepsilon_{kpj}\omega_p.$$

Therefore, we have, from (2.10),

$$\begin{aligned} \delta \int_D f_d &= \varepsilon_{kpj}\omega_p \int_D \left[\sigma^d_{kj} - \mathbf{h}_k\mathbf{n}_j + \beta\left(\mathbf{n}_j - \cos\alpha \frac{\partial_j\varphi}{|\nabla\varphi|}\right)\partial_k\varphi - \lambda\mathbf{n}_k\mathbf{n}_j - \beta\partial_k\varphi\mathbf{n}_j\right] d\mathbf{r} \\ &\quad + \omega_p \int_{\partial D}\left(\varepsilon_{kpq}\pi_{kj}\mathbf{n}_q\nu_j - \frac{\partial f}{\partial(\Delta\varphi)}\varepsilon_{kpj}(\partial_k\varphi)\nu_j\right) ds \\ &= \varepsilon_{kpj}\omega_p \int_D \left[\sigma^d_{kj} - \mathbf{h}_k\mathbf{n}_j\right] d\mathbf{r} \\ &\quad + \omega_p \int_{\partial D}\left(\varepsilon_{kpq}\pi_{kj}\mathbf{n}_q\nu_j - \frac{\partial f}{\partial(\Delta\varphi)}\varepsilon_{kpj}(\partial_k\varphi)\nu_j\right) ds = 0. \end{aligned} \tag{2.11}$$

From the equilibrium condition (2.8), we obtain

$$\int_D \varepsilon_{kpj}(\sigma^d_{kj})d\mathbf{r} + \int_{\partial D} \varepsilon_{kpq}\left(\mathbf{s}_k\mathbf{n}_q + (\partial_q\varphi)\,\mathbf{m}_k\right) ds = 0,$$

where $\mathbf{s}_i = \pi_{ij}\nu_j$ and $\mathbf{m}_i = \frac{\partial f}{\partial(\Delta\varphi)}\nu_i$. Using the notation Γ for the antisymmetric part of the elastic stress, we have

$$\varepsilon_{pji}\sigma^d_{ij} = \Gamma_p(\sigma^d), \tag{2.12}$$

and using the fact $\Gamma_p(\sigma^e) = \Gamma_p(\sigma^d)$, we have

$$\int_D \Gamma_p(\sigma^e)\,d\mathbf{r} + \int_{\partial D}(\mathbf{n}\times\mathbf{s} + \nabla\varphi\times\mathbf{m}) = 0. \tag{2.13}$$

From (2.9), we have

$$\begin{aligned} 0 &= \int_D \varepsilon_{jpq}\left(\partial_i(\sigma^e_{ji})\right)\mathbf{r}_q \\ &= \varepsilon_{jpq}\left(-\int_D(\sigma^e_{ji} + \phi_{ji})\partial_i\mathbf{r}_q + \int_{\partial D}(\sigma^e_{ji})\,\mathbf{r}_q\,\nu_i\right) \\ &= -\int_D \varepsilon_{jpi}(\sigma^e_{ji}) + \varepsilon_{jpq}\int_{\partial D}(\sigma^e_{ji})\,\mathbf{r}_q\,\nu_i. \end{aligned}$$

Therefore,

$$\int_D \Gamma_p(\sigma^e)\,d\mathbf{r} = \int_{\partial D}\mathbf{r}\times\mathbf{t}\,d\mathbf{x},$$

where $\mathbf{t}_j = (\sigma^e_{ji})\,\nu_i$. Inserting this equation into (2.13), we finally obtain

$$\int_{\partial D}(\mathbf{r}\times\mathbf{t} + \mathbf{n}\times\mathbf{s} + \nabla\varphi\times\mathbf{m})\,ds = 0. \tag{2.14}$$

This indicates that there are three contributions to these surface torques: mechanical torque (due to the stress tensor), director torque, and layer torque. This is the analogue of the balance of torques in nematic liquid crystals given by equation (3.115) of [8].

3. Hydrodynamic theory.

In this section, we derive the hydrodynamic equations for smectic C liquid crystals following previous work by Ericksen and Leslie (see [8], [4], [29], and [17]) for nematics and work by W. E [9] for smectic A. As we mentioned in the introduction, both the director and the layer functions are hydrodynamic variables.

3.1. Balance laws.

The equations of balance of mass, linear momentum, energy, and angular momentum are given by

$$\frac{d}{dt}\int_D \rho\,d\mathbf{x} = 0, \tag{3.1}$$

$$\frac{d}{dt}\int_D \rho\mathbf{v}_i\,d\mathbf{x} = \int_{\partial D}\sigma_{ij}ds_j, \tag{3.2}$$

$$\frac{d}{dt}\int_D E\,d\mathbf{x} = \int_{\partial D}(\sigma\mathbf{v} + \dot{\varphi}\tau + \dot{\mathbf{n}}\pi)\cdot\,ds - \int_{\partial D}\mathbf{q}\cdot ds, \tag{3.3}$$

$$\frac{d}{dt}\int_D(\mathbf{r}\times\rho\mathbf{v})\,d\mathbf{r} = \int_{\partial D}(\mathbf{r}\times\mathbf{t} + \mathbf{n}\times\mathbf{s} + \nabla\varphi\times\mathbf{m})\,ds, \tag{3.4}$$

where $D \subseteq \Omega$. We neglect body forces for simplicity, but they can easily be included as needed. Here, σ, τ, π, $\mathbf{t}$, $\mathbf{s}$, and $\mathbf{m}$ consist of the equilibrium components obtained in the previous sections and the dissipative components to be calculated next. The energy density in (3.3) is given by $E = \frac{1}{2}|\mathbf{v}|^2 + e$, where e denotes the internal energy per unit mass. The terms on the right-hand side of (3.3) represent the work done by the stress, the layer permeation force, and the director force, respectively, on the material. The vector field $\mathbf{q}$ denotes the heat flux.

Since the above balance laws are valid for any $D \subseteq \Omega$, by the Reynolds transport theorem, (3.1), (3.2), and (3.3) yield

$$\rho_t + \nabla \cdot (\rho \mathbf{v}) = 0 \quad \text{or} \quad \dot{\rho} = -\rho \nabla \cdot \mathbf{v}, \tag{3.5}$$

$$\rho \dot{\mathbf{v}} = \nabla \cdot \sigma \quad \text{or} \quad \rho \dot{\mathbf{v}}_i = \partial_j \sigma_{ij}, \tag{3.6}$$

$$\rho \left(\frac{1}{2}\mathbf{v}^2 + e\right)^{\cdot} = \nabla \cdot (\sigma \mathbf{v} + \dot{\varphi}\tau + \dot{\mathbf{n}}\pi) - \nabla \cdot \mathbf{q}, \tag{3.7}$$

where $\dot{f} = \frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f$ is a material derivative. The local form of (3.4) becomes a symmetry relation on constitutive equations, analogous to (2.28) of [10] for the nematic liquid crystal in the static case. It is guaranteed to hold for fields satisfying the balance of linear momentum, provided that the constitutive equations are invariant under rigid body rotations. It will be used later in the connection with the entropy inequality. In this paper, we assume incompressibility of the flow; that is, $\nabla \cdot \mathbf{v} = 0$ holds, and consequently, ρ is constant.

3.2. The entropy inequality. We assume that the second law of thermodynamics in the form of the Clausius–Duhem inequality

$$\rho \dot{S} + \nabla \cdot \left(\frac{\mathbf{q}}{T}\right) \geq 0 \tag{3.8}$$

holds for all processes. Here field S denotes the entropy of the system per unit mass. We use this inequality to determine the forms of the dissipative contribution to stresses and forces [17]. Taking (3.6) into account, we rewrite the balance of energy (3.7) as follows:

$$\rho \dot{e} = -\nabla \cdot \mathbf{q} + Tr(\sigma \nabla \mathbf{v}) + \nabla \cdot (\dot{\varphi}\tau + \dot{\mathbf{n}}\pi).$$

We let

$$H = e - TS \tag{3.9}$$

denote the Helmholtz free energy density. Substituting (3.9) into inequality (3.8), using the balance of energy, and omitting the pure divergence terms, we obtain

$$\begin{aligned} \rho \dot{H} &= \rho(\dot{e} - T\dot{S} - S\dot{T}) \\ &\leq Tr(\sigma \nabla \mathbf{v}) - \nabla \cdot \mathbf{q} + \nabla \cdot \left(\frac{\mathbf{q}}{T}\right) T - \rho S \dot{T} \\ &= Tr(\sigma \nabla \mathbf{v}) - \rho S \dot{T} - \frac{\mathbf{q} \cdot \nabla T}{T}. \end{aligned}$$

Since $H = H(\rho, \mathbf{n}, \nabla \mathbf{n}, \nabla \varphi, \Delta \varphi, T)$,

$$\dot{H} = \frac{\partial H}{\partial T}\dot{T} + \frac{\partial H}{\partial \mathbf{n}}\dot{\mathbf{n}} + \frac{\partial H}{\partial (\nabla \mathbf{n})}(\nabla \mathbf{n})^{\cdot} + \frac{\partial H}{\partial (\nabla \varphi)}(\nabla \varphi)^{\cdot} + \frac{\partial H}{\partial (\Delta \varphi)}(\Delta \varphi)^{\cdot} + \frac{\partial H}{\partial \rho}\dot{\rho}.$$

A direct computation gives

$$\begin{aligned}\nabla\dot{\varphi} &= (\nabla\varphi)^{\cdot} + \nabla\mathbf{v}\nabla\varphi,\\ \nabla\dot{\mathbf{n}} &= (\nabla\mathbf{n})^{\cdot} + \nabla\mathbf{v}\nabla\mathbf{n},\\ \Delta\dot{\varphi} &= (\Delta\varphi)^{\cdot} + 2Tr(\nabla^2\varphi\nabla\mathbf{v}) + \Delta\mathbf{v}\cdot\nabla\varphi.\end{aligned}$$

So, we get

$$\begin{aligned}&Tr(\sigma\nabla\mathbf{v}) - \frac{1}{T}\mathbf{q}\cdot\nabla T - \rho\dot{T}\left(S + \frac{\partial H}{\partial T}\right)\\ &\quad -\rho\frac{\partial H}{\partial\mathbf{n}}\dot{\mathbf{n}} - \rho\frac{\partial H}{\partial(\nabla\mathbf{n})}\{\nabla\dot{\mathbf{n}} - \nabla\mathbf{v}\nabla\mathbf{n}\} - \rho\frac{\partial H}{\partial(\nabla\varphi)}\{\nabla\dot{\varphi} - (\nabla\mathbf{v})(\nabla\varphi)\}\\ &\quad -\rho\frac{\partial H}{\partial(\Delta\varphi)}\left\{\Delta\dot{\varphi} - 2Tr(\nabla^2\varphi\nabla\mathbf{v}) - \Delta\mathbf{v}\cdot\nabla\varphi - \rho\frac{\partial H}{\partial\rho}\dot{\rho}\right\} \geq 0.\end{aligned}$$

Since, in particular, such an inequality holds for all possible choices of $\dot{T}$, we find that $S = -\frac{\partial H}{\partial T}$. For smectic C liquid crystals, we take $\tilde{f} = \rho H$, as in (2.2). Moreover, since the density is constant, the previous inequality becomes

$$\begin{aligned}&Tr\left[\left(\sigma + \frac{\partial\tilde{f}}{\partial(\nabla\mathbf{n})}\nabla\mathbf{n} + \frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\nabla^2\varphi - \nabla\left\{\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\nabla\varphi\right\}\right)\nabla\mathbf{v}\right] - \frac{1}{T}\mathbf{q}\cdot\nabla T\\ &-\left(\frac{\partial\tilde{f}}{\partial\mathbf{n}} - \nabla\frac{\partial\tilde{f}}{\partial(\nabla\mathbf{n})}\right)\dot{\mathbf{n}} - \left(-\nabla\cdot\frac{\partial\tilde{f}}{\partial(\nabla\varphi)} + \Delta\frac{\partial\tilde{f}}{\partial(\Delta\varphi)}\right)\dot{\varphi} \geq 0.\end{aligned}$$

Assuming that the stress consists of elastic and dissipative parts, σ^e (equation (2.6)) and σ^v, respectively, we write

$$\sigma = \sigma^e + \sigma^v.$$

This, together with substituting (2.3), (2.4), and (2.5) into the inequality, gives

$$Tr(\sigma^v\nabla\mathbf{v}) - \frac{1}{T}\mathbf{q}\cdot\nabla T + \mathbf{h}\cdot\dot{\mathbf{n}} + g\dot{\varphi} \geq 0.$$

Let us introduce the notation

$$\begin{aligned}2\sigma^{sym} &= \sigma^v + (\sigma^v)^T,\\ 2D &= \nabla\mathbf{v} + (\nabla\mathbf{v})^T, \qquad 2\mathbf{w} = \nabla\times\mathbf{v}.\end{aligned}$$

We denote $D = (d_{ij})$. From (2.12), we have

$$\Gamma(\sigma) = (\sigma_{32} - \sigma_{23}, \sigma_{13} - \sigma_{31}, \sigma_{21} - \sigma_{12}). \tag{3.10}$$

Therefore,

$$Tr(\sigma^{sym}D) + \Gamma(\sigma^v)\cdot\mathbf{w} + \mathbf{h}\cdot\dot{\mathbf{n}} + g\dot{\varphi} - \frac{1}{T}\mathbf{q}\cdot\nabla T \geq 0. \tag{3.11}$$

Following de Gennes [8], in order to characterize $\Gamma(\sigma^v)$, we need to consider the balance of angular momentum. Using integration by parts, (3.4) becomes

$$\int_D \mathbf{r}\times\rho\dot{\mathbf{v}}\,d\mathbf{r} = \int_D (\Gamma(\sigma) + \mathbf{r}\times(\nabla\cdot\sigma))\,d\mathbf{r} + \int_{\partial D} (\mathbf{n}\times\mathbf{s} + \nabla\varphi\times\mathbf{m})\,ds.$$

By substituting (3.6) into this, we get

$$\int_D \Gamma(\sigma)\,d\mathbf{x} + \int_{\partial D} (\mathbf{n}\times\mathbf{s} + \nabla\varphi\times\mathbf{m})\;ds = 0.$$

Moreover, using (2.11), we now get

$$\int_D \Gamma(\sigma^v)\,d\mathbf{r} = \int_D (-\mathbf{n}\times\mathbf{h})\,d\mathbf{r}, \tag{3.12}$$

which, substituted into the inequality, gives

$$Tr(\sigma^{sym}D) + \mathbf{h}\cdot\mathbf{N} + g\dot{\varphi} - \frac{1}{T}\mathbf{q}\cdot\nabla T \geq 0, \tag{3.13}$$

where $\mathbf{N} = \dot{\mathbf{n}} - \mathbf{w}\times\mathbf{n}$. Observe that $\mathbf{n}\cdot\mathbf{N} = 0$ holds as a result of the constraint $|\mathbf{n}| = 1$.

We conclude this subsection by writing the skew part of the viscous stress. Since (3.12) holds for all $D \subseteq \Omega$, we have

$$\sigma_{ij}^{skw} = \frac{1}{2}\varepsilon_{kji}\Gamma_k = -\frac{1}{2}\varepsilon_{kji}\mathbf{n}_j h_i. \tag{3.14}$$

From now on, we will denote $\mathbf{l} = \nabla\varphi$.

3.3. Coefficients of viscosity. We consider linear dependence of the dissipative forces on their fluxes. We require $\mathbf{q}$ and σ^{sym} to be invariant under the simultaneous transformations $\mathbf{n}\to-\mathbf{n}$ and $\nabla\varphi\to-\nabla\varphi$. We impose that $\mathbf{h}$ and $\dot{\varphi}$ change into $-\mathbf{h}$ and $-\dot{\varphi}$, respectively, under the same transformation. Hence the most general form of the equations is

$$\sigma_{ij}^{sym} = A_{ijk}^1\mathbf{N}_k + A_{ijkm}^2 d_{km}, \tag{3.15}$$

$$\mathbf{h}_i = B_{ij}^1\mathbf{N}_j + B_{ijk}^2 d_{jk}, \tag{3.16}$$

$$\mathbf{q}_i = C_{ij}^3\frac{1}{T}\frac{\partial T}{\partial x_j} + C_i^4 g, \tag{3.17}$$

$$\dot{\varphi} = D_i^3\frac{1}{T}\frac{\partial T}{\partial x_i} + D^4 g, \tag{3.18}$$

where A, B, C, and D are functions of $\mathbf{n}_i$ and $\partial_i\varphi$. The most general form of $\mathbf{h}$ that meets the invariance requirement is

$$\begin{aligned}\mathbf{h} = {}&\beta_1\mathbf{N} + \beta_2(\mathbf{N}\cdot\mathbf{l})\mathbf{l} + \beta_3 D\mathbf{n} + \beta_4 D\mathbf{l} + \beta_5(\mathbf{n}\cdot D\mathbf{l})\mathbf{l}\\ &+ \beta_6(\mathbf{n}\cdot D\mathbf{n})\mathbf{l} + \beta_7(\mathbf{l}\cdot D\mathbf{l})\mathbf{l}.\end{aligned} \tag{3.19}$$

The skew-symmetric part of the viscous stress follows from (3.14) together with (3.19). Writing the symmetric part of the stress tensor explicitly from (3.15) and adding to it the skew part, gives

$$\begin{aligned}\sigma^v = {}&\alpha_1 D + \alpha_2 D\mathbf{n}\otimes\mathbf{n} + \alpha_3\mathbf{n}\otimes D\mathbf{n} + \alpha_4(D\mathbf{l}\otimes\mathbf{l} + \mathbf{l}\otimes D\mathbf{l}) + \alpha_5 D\mathbf{l}\otimes\mathbf{n}\\ &+ \alpha_6\mathbf{n}\otimes D\mathbf{l} + \alpha_7(D\mathbf{n}\otimes\mathbf{l} + \mathbf{l}\otimes D\mathbf{n}) + \alpha_8(\mathbf{l}\cdot D\mathbf{l})\mathbf{l}\otimes\mathbf{n} + \alpha_9(\mathbf{l}\cdot D\mathbf{l})\mathbf{n}\otimes\mathbf{l}\\ &+ \alpha_{10}(\mathbf{n}\cdot D\mathbf{l})\mathbf{l}\otimes\mathbf{l} + \alpha_{11}(\mathbf{l}\cdot D\mathbf{l})\mathbf{n}\otimes\mathbf{n} + \alpha_{12}(\mathbf{n}\cdot D\mathbf{l})\mathbf{l}\otimes\mathbf{n}\\ &+ \alpha_{13}(\mathbf{n}\cdot D\mathbf{l})\mathbf{n}\otimes\mathbf{l} + \alpha_{14}(\mathbf{n}\cdot D\mathbf{n})\mathbf{l}\otimes\mathbf{l} + \alpha_{15}(\mathbf{n}\cdot D\mathbf{l})\mathbf{n}\otimes\mathbf{n}\\ &+ \alpha_{16}(\mathbf{n}\cdot D\mathbf{n})\mathbf{l}\otimes\mathbf{n} + \alpha_{17}(\mathbf{n}\cdot D\mathbf{n})\mathbf{n}\otimes\mathbf{l} + \alpha_{18}(\mathbf{n}\cdot D\mathbf{n})\mathbf{n}\otimes\mathbf{n}\\ &+ \alpha_{19}(\mathbf{l}\cdot D\mathbf{l})\mathbf{l}\otimes\mathbf{l} + \alpha_{20}(\mathbf{l}\otimes\mathbf{N} + \mathbf{N}\otimes\mathbf{l}) + \alpha_{21}\mathbf{N}\otimes\mathbf{n} + \alpha_{22}\mathbf{n}\otimes\mathbf{N}\\ &+ \alpha_{23}(\mathbf{l}\cdot\mathbf{N})\mathbf{l}\otimes\mathbf{l} + \alpha_{24}(\mathbf{l}\cdot\mathbf{N})\mathbf{l}\otimes\mathbf{n} + \alpha_{25}(\mathbf{l}\cdot\mathbf{N})\mathbf{n}\otimes\mathbf{l} + \alpha_{26}(\mathbf{l}\cdot\mathbf{N})\mathbf{n}\otimes\mathbf{n}.\end{aligned} \tag{3.20}$$

We also get

$$
\begin{aligned}
\mathbf{q} &= \frac{1}{T}(\mu_1\partial_i T + \mu_2\mathbf{n}_i\mathbf{n}_j\partial_j T + \mu_3\varphi_i\varphi_j\partial_j T + \mu_4\mathbf{n}_i\varphi_j\partial_j T + \mu_5\mathbf{n}_j\varphi_i\partial_j T)\\
&\quad + (\gamma_1\mathbf{n}_i + \gamma_2\varphi_i)g,\\
\dot\varphi &= \frac{1}{T}(\gamma_1'\mathbf{n}_i\partial_i T + \gamma_2'\varphi\partial_i T) + \gamma_3' g.
\end{aligned}
\tag{3.21}
$$

The viscosity coefficients α, β, γ, μ, and γ' cannot be arbitrarily chosen: they satisfy inequalities that follow from (3.13). Further restrictions result from Onsager's reciprocal relations (see, for instance, [8], [4], and [22]):

$$
\begin{aligned}
&\beta_1 = \alpha_{22} - \alpha_{21}, \qquad \beta_2 = \alpha_{25} - \alpha_{24}, \qquad \beta_3 = \alpha_3 - \alpha_2 = \alpha_{21} + \alpha_{22},\\
&\beta_4 = \alpha_6 - \alpha_5 = 2\alpha_{20}, \qquad \beta_5 = \alpha_{13} - \alpha_{12} = \alpha_{24} + \alpha_{25},\\
&\beta_6 = \alpha_{17} - \alpha_{16} = \alpha_{26}, \qquad \beta_7 = \alpha_9 - \alpha_8 = \alpha_{23},\\
&\gamma_1' = \gamma_1, \qquad \gamma_2' = \gamma_2.
\end{aligned}
\tag{3.22}
$$

We end this section by summarizing the governing equations of the hydrodynamic of smectic C. It consists of 11 equations and 11 unknowns. The latter are the pressure p, the velocity field $\mathbf{v}$, the director $\mathbf{n}$, the layer φ, the temperature T, and Lagrange multipliers λ and β. The equations are as follows:

- balance of linear momentum equation (3.6):

$$\rho\dot{\mathbf{v}} = \nabla\cdot(-pI + \sigma^d + \sigma^v) + \mathbf{f}, \tag{3.23}$$

  where the elastic stress σ^d is given from (2.5), the viscous part σ^v is from (3.20), and $\mathbf{f}$ is an external force;
- molecular field equation:

$$\frac{\partial f_d}{\partial \mathbf{n}_i} - \partial_j\pi_{ij} + \lambda\mathbf{n}_i + \beta\partial_i\varphi + \mathbf{h}_i = 0, \tag{3.24}$$

  where $\mathbf{h}$ is given from (3.19);
- permeation force equation:

$$\dot\varphi = \frac{1}{T}(\gamma_1'\mathbf{n}_i\partial_i T + \gamma_2'\varphi\partial_i T) + \gamma_3' g,$$

  where g is defined in (2.3);
- balance of energy equation:

$$E_t + \nabla\cdot(E\mathbf{v} + \mathbf{q} - \sigma\mathbf{v} - \dot\varphi\tau - \dot{\mathbf{n}}\pi) = 0,$$

  where $\mathbf{q}$ is given from (3.21);
- incompressibility condition:

$$\nabla\cdot\mathbf{v} = 0;$$

- two constraints:

$$|\mathbf{n}| = 1 \qquad \text{and} \qquad \mathbf{n}\cdot\nabla\varphi = \cos\alpha|\nabla\varphi|.$$

In the isothermal case, this reduces to 10 equations and 10 unknowns.

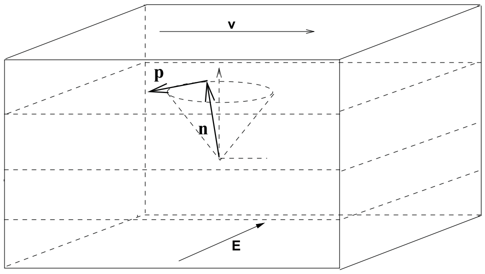


FIG. 1. *Smectic C* liquid crystals in the homeotropic geometry; the polarization is perpendicular to both the layer normal and the director.*

4. Switching dynamics.

4.1. Electrostatic energy. In order to investigate the electric effect, we consider the electric energy [8], [29],

$$f_e = -\int_\Omega \mathbf{D}\cdot d\mathbf{E} = -\int_\Omega (\varepsilon_\perp \mathbf{E} + \varepsilon_a(\mathbf{n}\cdot\mathbf{E})\mathbf{n} + \mathbf{P})\cdot d\mathbf{E}, \tag{4.1}$$

where $\mathbf{E}$ denotes the electric field, $\mathbf{P}$ denotes the ferroelectric polarization, and ε_a represents the dielectric anisotropy. Since chiral smectic C liquid crystals are known to be ferroelectric, they possess a spontaneous polarization $\mathbf{P}$. Dropping the constant term in the electric energy, (4.1) reduces to

$$f_e = -\frac{1}{2}\int_\Omega \varepsilon_a(\mathbf{n}\cdot\mathbf{E})^2\,d\mathbf{x} - \int_\Omega \mathbf{P}\cdot\mathbf{E}\,d\mathbf{x}. \tag{4.2}$$

Since the magnitude of the polarization is small in smectic C* liquid crystals, we assume that it is constant, P_0. Furthermore, since the chiral molecules create a spontaneous polarization within each layer and the polarization is perpendicular to the director (see Figure 1), we write

$$\mathbf{P} = P_0\frac{\nabla\varphi\times\mathbf{n}}{|\nabla\varphi\times\mathbf{n}|}. \tag{4.3}$$

The polarization $\mathbf{P}$ also gives an electrostatic charge density, $-\nabla\cdot\mathbf{P}$. The electrostatic effects of chiral smectic C liquid crystals were studied in [24]. As a result of the electric field, the Lorentz force $\mathbf{f} = (-\nabla\cdot\mathbf{P})\mathbf{E}$ has to be included in (3.23).

4.2. The model. We consider the homeotropic geometry where the liquid crystal is confined between two parallel plates with the smectic layers parallel to the plates (Figure 1). Let

$$\begin{aligned} \mathbf{n} &= (\cos\phi\sin\alpha, \sin\phi\sin\alpha, \cos\alpha),\\ \mathbf{v} &= (v,0,0),\\ \nabla\varphi &= (0,0,k), \end{aligned} \tag{4.4}$$

with $\phi = \phi(z)$, $v = v(z)$, $p = p(x,y,z)$, and α and k constant. We consider the switching dynamics between states with opposite polarization when a uniform electric field is applied in a direction parallel to the layer, i.e., $\mathbf{E} = E_0(0,1,0)$ in (4.2). We also restrict our attention to the case when $\varepsilon_a < 0$, which applies to many smectic C liquid crystals. This tends to align the director and polarization fields along directions perpendicular and parallel to the applied field, respectively. Note that, since $\nabla\cdot\mathbf{P} = 0$, $\mathbf{f}$ in (3.23) vanishes.

The balance of linear momentum (3.23) yields

$$\begin{aligned} \rho v_t &= -\frac{\partial p}{\partial x} + \frac{\partial}{\partial z}(\sigma^d_{13} + \sigma^v_{13}), \\ 0 &= -\frac{\partial p}{\partial y} + \frac{\partial}{\partial z}(\sigma^d_{23} + \sigma^v_{23}), \\ 0 &= -\frac{\partial p}{\partial z} + \frac{\partial}{\partial z}(\sigma^d_{33} + \sigma^v_{33}). \end{aligned} \tag{4.5}$$

The first two equations imply that p is linear on x and y. So, p is of the form

$$p(x,y,z,t) = k_0(t) + k_1(t)x + k_2(t)y + \sigma^d_{33} + \sigma^v_{33}.$$

Hence, (4.5) reduces to

$$\rho v_t = -k_1(t) + \frac{\partial}{\partial z}(\sigma^d_{13} + \sigma^v_{13}).$$

Also, as a result of (4.4), equation (3.24) reduces to a single equation for ϕ. Hence the system of governing equations is

$$\begin{aligned} \rho v_t &= \frac{\partial}{\partial z}\Big[g(\phi)v_z - \eta_3(\sin\phi)\phi_t\Big] - k_1(t), \\ 2\beta_1 \sin^2\alpha\phi_t &= \lambda_1 \sin\phi v_z + 2K\phi_{zz} - 2P_0E_0\sin\phi - |\varepsilon_a|E_0^2\sin^2\alpha\sin 2\phi, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} &\lambda_1 = \sin\alpha((-\beta_1+\beta_3)\cos\alpha + \beta_4 k), \\ &K = \sin^2\alpha(K_2\sin^2\alpha + K_3\cos^2\alpha), \\ &g(\phi) = \frac{1}{2}(\eta_1 + \eta_2\cos^2\phi), \\ &\eta_1 = \alpha_1 + \alpha_4 k^2(\alpha_2 - \alpha_{21})\cos^2\alpha + (\alpha_5 + \alpha_7 - \alpha_{20})k\cos\alpha, \\ &\eta_2 = \sin^2\alpha\Big(\alpha_3 + \alpha_{22} + k^2(\alpha_{13} + \alpha_{25}) + (\alpha_{15} + 2\alpha_{17} + \alpha_{26})k\cos\alpha \\ &\qquad\qquad +2\alpha_{18}\cos^2\alpha\Big), \\ &\eta_3 = \sin\alpha(\alpha_{20}\,k + \alpha_{21}\cos\alpha), \\ &\frac{\partial p}{\partial x} = k_1(t). \end{aligned}$$

Onsager reciprocal relation (3.22) now gives

$$\lambda_1 = 2\sin\alpha\,(\alpha_{20}k + \alpha_{21}\cos\alpha) = 2\eta_3. \tag{4.7}$$

Note that we may derive the following inequalities from the dissipation inequality (3.13):

$$\beta_1 > 0, \quad g(\phi) > 0, \quad \text{and} \quad \beta_1 g(\phi)\sin^2\alpha - \eta_3^2\sin^2\phi > 0. \tag{4.8}$$

The first inequality can be obtained from the shear flow alignment.

Let $V = v + K(t)/\rho$, where $K(t)$ is an antiderivative of $k_1(t)$. Using (4.7) and new variables

$$\bar{z} = \left(\frac{P_0 E_c}{K}\right)^{\frac{1}{2}} z, \quad \bar{t} = \frac{P_0 E_c}{\beta_1 \sin^2\alpha} t, \quad u = \frac{\beta_1}{\sqrt{K P_0 E_c}} V, \quad \text{and} \quad E_c = \frac{2P_0}{|\varepsilon_a| \sin^2\alpha},$$

the system (4.6) becomes

$$\begin{aligned} \varepsilon u_{\bar{t}} &= \frac{\partial}{\partial \bar{z}} \left(\frac{\sin^2\alpha}{\eta_1} g(\phi)\, u_{\bar{z}} - \frac{\eta_3}{\eta_1} (\sin\phi) \phi_{\bar{t}} \right), \\ \phi_{\bar{t}} &= \frac{\eta_3}{\beta_1} \sin\phi\, u_{\bar{z}} + \phi_{\bar{z}\bar{z}} - e \sin\phi - e^2 \sin 2\phi, \end{aligned} \tag{4.9}$$

where

$$\varepsilon = \frac{\rho K}{\beta_1 \eta_1} \qquad \text{and} \qquad e = E_0/E_c.$$

We may assume that the dimensionless parameter $\varepsilon \ll 1$ since the viscous coefficients are much bigger than the elastic coefficients.

4.3. Traveling wave solution. In this section, we study the traveling wave solutions of system (4.9) to understand the director profile when the electric field is applied. For this, we look for a solution of (4.9) in the form $w(\zeta) = u(\bar{z}, \bar{t})$ and $\theta(\zeta) = \phi(\bar{z}, \bar{t})$, where $\zeta = \bar{z} - c\bar{t}$, such that θ connects two bistable states, $\theta = 0$ and $\theta = \pi$. Then the traveling wave solution is $(w(\zeta), \theta(\zeta)) \in C^2(\mathbb{R}) \times C^2(\mathbb{R})$ and $c \in \mathbb{R}$ satisfying

$$-c\varepsilon\, w' = \left[\frac{\sin^2\alpha}{\eta_1}\, g(\theta) w' + c\frac{\eta_3}{\eta_1} \sin\theta\, \theta' \right]', \tag{4.10}$$

$$-c\, \theta' = \frac{\eta_3}{\beta_1} \sin\theta w' + \theta'' - e\sin\theta - e^2 \sin 2\theta, \tag{4.11}$$

with $w(-\infty) = w'(-\infty) = 0$, $\theta(-\infty) = 0$, and $\theta(\infty) = \pi$. This is consistent with the system approaching an equilibrium state. Here the $'$ denotes the derivative with respect to ζ. Integrating (4.10) and using the relations $w(-\infty) = w'(-\infty) = \theta(-\infty) = 0$, equation (4.10) reduces to

$$-c\varepsilon\, w = \frac{\sin^2\alpha}{\eta_1}\, g(\theta) w' + c\frac{\eta_3}{\eta_1} \sin\theta\, \theta'. \tag{4.12}$$

Substituting this into (4.11), we have

$$\theta'' + ch(\theta)\theta' - e\sin\theta - e^2 \sin 2\theta - c\varepsilon \frac{\eta_3 \sin\theta}{\beta_1 \sin^2\alpha\, g(\theta)} w = 0, \tag{4.13}$$

where

$$0 < h(\theta) := 1 - \frac{\eta_3^2}{\beta_1 \sin^2\alpha} \frac{\sin^2\theta}{g(\theta)} \leq 1. \tag{4.14}$$

Notice that the inequalities follow from (4.8). Introducing the rescaled variable $v = \varepsilon w$, we rewrite the system (4.12) and (4.13) as follows:

$$\begin{aligned} &v' + c\,\varepsilon b(\theta) v + c\,\varepsilon \beta_1 d(\theta) \theta' = 0, \\ &\theta'' + ch(\theta)\theta' + E(\theta) - c\, d(\theta) v = 0, \end{aligned} \tag{4.15}$$

where

$$\begin{aligned} b(\theta) &= \frac{\eta_1}{\sin^2\alpha\, g(\theta)}, \\ E(\theta) &= -e\sin\theta - e^2\sin 2\theta, \\ d(\theta) &= \frac{\eta_3}{\beta_1\sin^2\alpha}\frac{\sin\theta}{g(\theta)}. \end{aligned}$$

If $\theta(\zeta)$ and $v(\zeta)$ are solutions of the system, then so are $\theta(\zeta+\zeta_0)$ and $v(\zeta+\zeta_0)$ for any constant ζ_0. Hence we impose a normalized condition, $\theta(0) = \frac{1}{2}[\theta(-\infty)+\theta(\infty)] = \frac{\pi}{2}$. Using the condition $v(-\infty) = 0$, we solve the first equation for v,

$$v(\varepsilon, c, \theta, \zeta) = -c\varepsilon\beta_1 e^{-c\varepsilon\beta(\zeta)}\int_{-\infty}^{\zeta} d(\theta(s))\theta'(s)e^{c\varepsilon\beta(s)}\,ds,$$

where

$$\beta(\zeta) = \int b(\theta(s))\,ds.$$

Substituting this expression into the second equation of (4.15), we have

$$\theta'' + ch(\theta)\theta' + E(\theta) - c\,d(\theta)v(\varepsilon, c, \theta, \zeta) = 0. \tag{4.16}$$

For $\varepsilon = 0$, equation (4.16) becomes

$$\theta'' + ch(\theta)\theta' + E(\theta) = 0. \tag{4.17}$$

From now on, we will restrict our attention to the case $e > \frac{1}{2}$ so that the term $E(\theta)$ is cubic-like. In fact, if $e > \frac{1}{2}$, the term $E(\theta) = -e\sin\theta(1+2e\cos\theta)$ has an intermediate zero. In this case, (4.17) with a bistable nonlinearity has an increasing traveling wave solution (c_0, θ_0) with $c_0 > 0$ that $\theta_0 \to 0$ as $\zeta \to -\infty$ and $\theta_0 \to \pi$ as $\zeta \to \infty$, thanks to the condition $h(\theta) > 0$ for any θ [15]. Furthermore, we can easily see that [11], [1], from the phase plane analysis, θ_0 satisfies

$$|\theta_0(\zeta) - \pi| \le Ke^{-\mu_1\zeta}, \qquad |\theta_0'(\zeta)| \le Ke^{-\mu_1\zeta}$$

for $\zeta \ge 0$ and for some constant $\mu_1 > 0$, and

$$|\theta_0(\zeta)| \le Ke^{\mu_2\zeta}, \qquad |\theta_0'(\zeta)| \le Ke^{\mu_2\zeta}$$

for $\zeta \le 0$ and for some constant $\mu_2 > 0$.

Motivated by the work in [12], we look for solutions of (4.16) of the form

$$\begin{aligned} \theta &= \theta_0 + s(\zeta, c, \varepsilon), \\ c &= c_0 + \sigma. \end{aligned}$$

Substituting these into (4.16) and letting $r = (s, \sigma)$, we define the operator F:

$$\begin{aligned} F(r;\varepsilon) = \theta_0'' + s'' + (c_0+\sigma)h(\theta_0+s)(\theta_0'+s') + E(\theta_0+s) \\ -(c_0+\sigma)d(\theta_0+s)v(\varepsilon, c_0+\sigma, \theta_0+s). \end{aligned}$$

Note that r satisfies the boundary conditions

$$r(-\infty;\varepsilon) = r(\infty;\varepsilon) = 0. \tag{4.18}$$

For a fixed constant μ satisfying $0 < \mu < \min\{\mu_1, \mu_2\}$, we define the function spaces

$$B^n_\mu(\mathbb{R}) = \left\{ u \in C^n(\mathbb{R}) : \|u\|_{B^n_\mu(\mathbb{R})} \equiv \sum_{i=0}^{n} \sup_{x\in\mathbb{R}} \left| e^{\mu|x|} \left(\frac{d}{dx}\right)^i u(x) \right| < \infty \right\},$$
$$\dot{B}^n_\mu(\mathbb{R}) = \{u \in B^n_\mu(\mathbb{R}) : u(0) = 0\}.$$

Note that the mapping F is differentiable from X into Y, where

$$X = \dot{B}^2_\mu \times \mathbb{R},$$
$$Y = B^0_\mu.$$

The next lemma establishes that F meets the hypotheses of the implicit function theorem.

LEMMA 4.1.

(i) *F is a continuous mapping, and $\|F(r;\varepsilon) - F(r;0)\|_Y \to 0$ as $\varepsilon \to 0$.*

(ii) *F is continuously Fréchet differentiable with respect to r, and*

$$\|F_r(r;\varepsilon)[\tilde{r}] - F_r(r;0)[\tilde{r}]\|_Y \to 0 \text{ as } \varepsilon \to 0.$$

(iii) *$F_r(0;0)$ has a bounded inverse.*

Proof.

(i) We have

$$\begin{aligned}
&|e^{\mu|z|}(F(r;\varepsilon) - F(r;0))| \\
&= e^{\mu|z|} \left| (c_0+\sigma)^2 d(\theta_0+s)\varepsilon\beta_1 g_-(z) \int_{-\infty}^{z} d(\theta_0+s)(\theta_0'+s')g_+(t)e^{\mu|t|}e^{-\mu|t|}\,dt \right| \\
&\le C\varepsilon(c_0+\sigma)\|\theta_0'+s'\|_{B^0_\mu} e^{\mu|z|} \left| \int_{-\infty}^{z} e^{-\mu|t|}dt \right| \\
&\le C\varepsilon(c_0+\sigma)\|\theta_0+s\|_{B^2_\mu},
\end{aligned}$$

where $g_\pm(z) = \exp(\pm\varepsilon(c_0+\sigma)\beta(z))$ and $\beta'(z) = b(\theta_0+s)(z)$. Notice that we used the fact that $\beta(z)$ is increasing.

(ii) Note that

$$\begin{aligned}
&e^{\mu|z|}|F_r(r;\varepsilon)\tilde{r} - F_r(r;0)\tilde{r}| \\
&= e^{\mu|z|}|\tilde{\sigma}d(\theta_0+s)v_\varepsilon(z) - (c_0+\sigma)d'(\theta_0+s)v_\varepsilon(z)\tilde{s} + (c_0+\sigma)d(\theta_0+s)v_\varepsilon'(z)\tilde{s}|,
\end{aligned}$$

where $v_\varepsilon(z) = v(\varepsilon, c_0+\sigma, \theta_0+s, z)$. Since we have

$$v_\varepsilon'(z) = -(c_0+\sigma)\varepsilon b(\theta_0+s)v_\varepsilon(z) - (c_0+\sigma)\varepsilon\beta_1 d(\theta_0+s)(\theta_0'+s'),$$

the rest of the proof follows as in part (i).

(iii) It suffices to show that for any $g \in Y$, the linear problem

$$F_r(0;0)\tilde{r} = g$$

has a unique solution $\tilde{r} \in \dot{B}^2_\mu$ such that

$$\|\tilde{r}\|_{\dot{B}^2_\mu} \le C\|g\|_Y.$$

The above linear problem can be explicitly written as

$$\tilde{s}^{''} + c_0 h(\theta_0)\tilde{s}' + (c_0 h(\theta_0)\theta_0' + E'(\theta_0))\tilde{s} = G, \tag{4.19}$$

where

$$G = g - h(\theta_0)\theta_0'\tilde{\sigma}.$$

The proof of existence and uniqueness of solution of (4.19) satisfying boundary condition (4.18) follows as that of Lemma 3 in [15]. □

The proof of the following theorem uses the implicit function theorem together with Lemma 4.1.

THEOREM 4.2. *For $\varepsilon > 0$ sufficiently small, there exists a unique (up to translation in ζ) $(c_\varepsilon, \theta_\varepsilon, v_\varepsilon)$ satisfying* (4.15) *such that*

$$\|\theta_\varepsilon - \theta_0\|_{B^2_\mu} + \|v_\varepsilon\|_{B^1_\mu} + |c_\varepsilon - c_0| \longrightarrow 0 \qquad \text{as } \varepsilon \to 0.$$

4.4. Speed of the traveling wave. In this section, we study the speed of the traveling front of (4.16). We follow the variational approach in [2] for reaction-diffusion equations.

We first consider front propagation for the reaction-diffusion equation

$$\theta'' + c\theta' = H(\theta), \tag{4.20}$$

where $H(\theta) = e \sin\theta + e^2 \sin 2\theta$. This is the traveling wave equation for a switching problem for smectic C liquid crystals, when flow effects are neglected (see [25], [26]). It is known that there exists a unique heteroclinic solution (c_s, θ_s) of (4.20) such that $\theta(-\infty) = 0$, $\theta(\infty) = \pi$, and $\theta'(\zeta) > 0$ for $|\zeta| < \infty$. This equation can also be obtained from our model by neglecting the flow. The authors of [7] found the explicit wave front solution (θ_s, c_s) of (4.20), given by

$$\theta_s(\zeta) = 2\arctan(e^{\sqrt{2}e\zeta}), \quad c_s = \frac{1}{\sqrt{2}}.$$

Following the work by Benguria and Depassier in [2], we obtain the variational expression of the speed for (4.20):

$$c_s^2 = \max \frac{2\int_0^\pi Hf\,d\theta}{\int_0^\pi \frac{f^2}{f'}d\theta}, \tag{4.21}$$

where the maximum is taken over all positive increasing functions f in $(0, \pi)$. We denote the maximizing function by $\hat{f}$.

With the help of Theorem 4.2, we will investigate c_0, the speed of the traveling wave solution of (4.17). We let θ be a solution of (4.17). The same proof as in [2] leads to the variational principle for the speed of the traveling wave of (4.17). Noting that $0 < h(\theta) \leq 1$,

$$c_0^2 = \max \frac{2\int_0^\pi Hf\,d\theta}{\int_0^\pi \frac{h^2 f^2}{f'}d\theta}, \tag{4.22}$$

where the maximum is taken over all positive increasing functions f in $(0, \pi)$ for which the integrals exist.

Now we compare the speeds c_0 and c_s using (4.21) and (4.22). From (4.22), we get

$$c_0^2 \geq \frac{2\int_0^\pi H\hat{f}d\theta}{\int_0^\pi \frac{h^2\hat{f}^2}{\hat{f}'}d\theta}.$$

Since $0 < h(\theta) \leq 1$, we have

$$h^2(\theta) \leq h(\theta) = 1 - \frac{2\eta_3^2 \sin^2\theta}{\beta_1 \sin^2\alpha\,(\eta_1 + \eta_2\cos^2\theta)} \leq 1 - A\sin^2\theta,$$

where

$$A := \frac{2\eta_3^2}{\beta_1 \sin^2\alpha\,(\eta_1 + \max\{\eta_2, 0\})} \geq 0. \tag{4.23}$$

It follows from (4.14) that $0 \leq A < 1$. In fact, if $\max\{\eta_2, 0\} = 0$, then

$$A = \frac{2\eta_3^2}{\beta_1\eta_1\sin^2\alpha} = 1 - h\left(\frac{\pi}{2}\right) < 1.$$

Also, if $\max\{\eta_2, 0\} = \eta_2$, then

$$A = \frac{2\eta_3^2}{\beta_1\sin^2\alpha(\eta_1 + \eta_2)} \leq 1 - h\left(\frac{\pi}{2}\right) < 1.$$

From (4.21) and (4.22), we have

$$c_0^2 \geq \frac{2\int_0^\pi H\hat{f}d\theta}{\int_0^\pi \frac{\hat{f}^2}{\hat{f}'}d\theta - A\int_0^\pi \frac{\sin^2\theta\,\hat{f}^2}{\hat{f}'}\,d\theta} = \frac{2\int_0^\pi H\hat{f}d\theta}{\int_0^\pi \frac{\hat{f}^2}{\hat{f}'}d\theta(1 - A\cdot M)} = \frac{c_s^2}{1 - A\cdot M}, \tag{4.24}$$

where M is a fixed number given by

$$M = \frac{\int_0^\pi \frac{\sin^2\theta\,\hat{f}^2}{\hat{f}'}\,d\theta}{\int_0^\pi \frac{\hat{f}^2}{\hat{f}'}d\theta}.$$

Notice that $0 < M < 1$ is independent of viscosity coefficients, since $\hat{f}$ is the maximizing function for c_s. We rewrite (4.24) as

$$\left(\frac{c_0}{c_s}\right)^2 \geq \frac{1}{1 - A\cdot M}. \tag{4.25}$$

The inequality (4.25) shows that the switching is faster when the flow is taken into consideration. The value A in (4.23) is the control parameter; i.e., A is the quantity which measures flow effects. If A is close to 1, then flow effects are expected to be strong, and if A is close to 0, then flow effects are weak. In particular, we see that the ratio of c_0 to c_s increases as A approaches 1. In view of this control parameter, the optimal switching time is obtained when $A = 1$. The sufficient condition for this is

$$\eta_2 \leq 0 \qquad \text{and} \qquad 2\eta_3^2 = \beta_1\eta_1\sin^2\alpha.$$

This condition depends only on the viscosity, the tilt angle, and the layer thickness of the material. In [5], the control parameter was also found. Our control parameter A is analogous to that given by Carlsson, Clark, and Zou in [5].

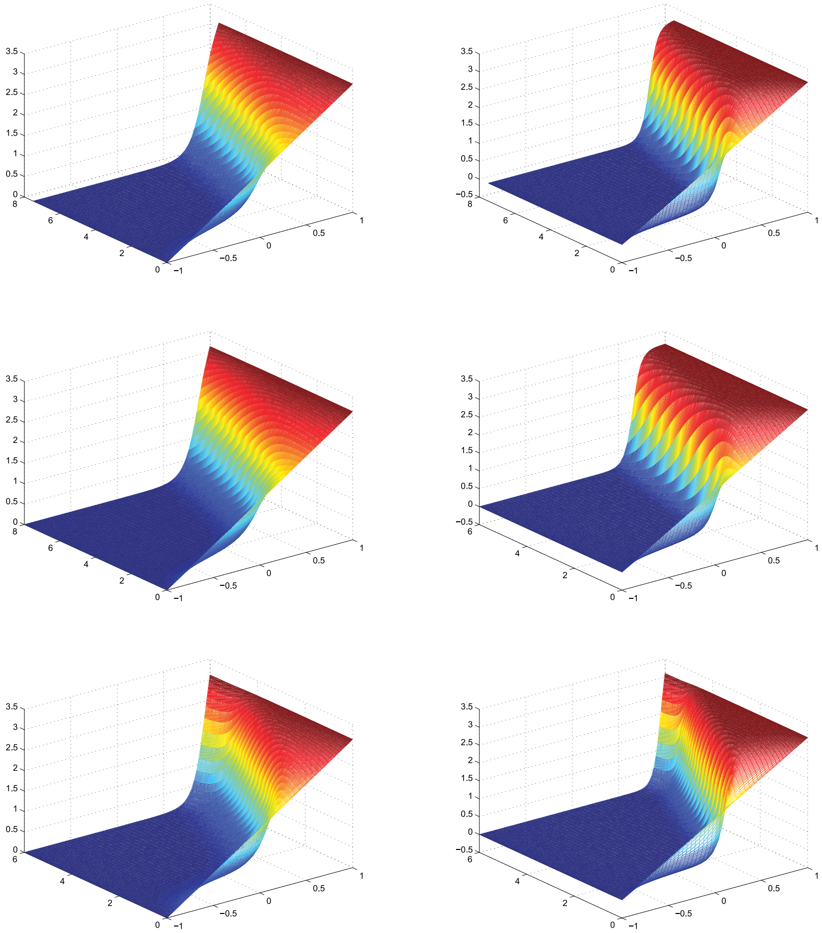


FIG. 2. *Director configuration when the initial condition is the linear function connecting* 0 *and* π. *The applied electric field corresponds to* $e = 0.75$ *on the left column and* $e = 1.3$ *on the right column. The upper row depicts the director profile when flow is neglected, while in the second and third rows the flow effects are included. For simulations in the second row, the control parameter* A *is close to* 0, *while for the third row,* A *is close to* 1.

4.5. Numerical simulation. In order to solve the system (4.9) numerically, we use a second order semi-implicit scheme for time discretization. This scheme requires us to solve two Helmholtz equations at each time step, which we do by means of a spectral Galerkin method (see [27] and [28]). We impose the homogeneous boundary and initial conditions on u and assume strong anchoring conditions for ϕ, i.e., $\phi(0,t) = 0$ and $\phi(L,t) = \pi$, where L is the domain size. From (4.3) and (4.4) we see that the director configurations $\phi = 0$ and $\phi = \pi$ correspond to the polarization pointing in the same and opposite directions as the applied electric field, respectively.

For ε, we simply take $\varepsilon = 10^{-6}$. For the tilt angle and viscosity coefficients appearing in the system, we use $\eta_1 = 3.8$, $\eta_2 = -0.2$, $\beta_1 = 40.9706$, and $\alpha = \pi/8$. This set of parameters, employed in [3], gives a value of the control parameter A in (4.23) of approximately 0.5. In [3], the authors study the macroscopic equations of

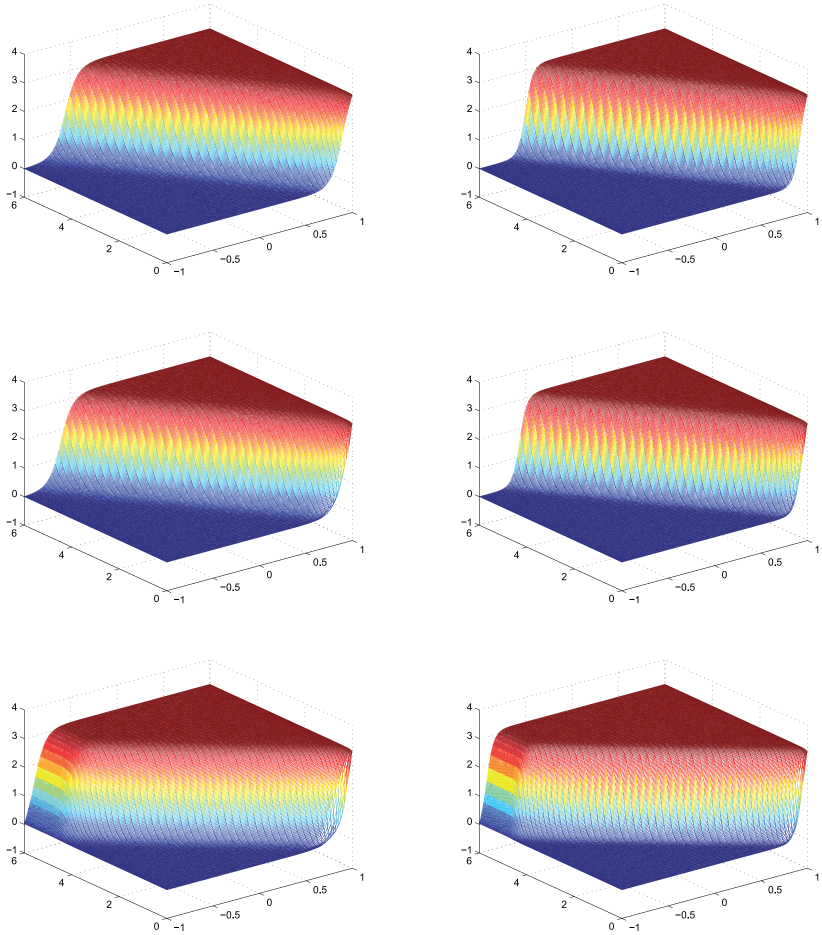


FIG. 3. *Director configuration when the sign of the electric field is reversed at $t = 0$. The arrangement of the simulations in rows and columns follows the analogous criteria to those in Figure* 2.

smectic C* liquid crystals [18] in a homeotropic geometry to investigate the backflow effect upon the removal of a strong electric field. Their approach is based on linear analysis, replacing the nonlinear functions by their initial values. In our simulations, we vary the parameter η_3 in order for A to span the interval $(0, 1)$.

We consider two types of initial conditions for ϕ corresponding to the simulations shown in Figures 2 and 3, respectively. In Figure 2, the initial value ϕ_0 is the linear function connecting two bistable states, 0 and π. When a positive electric field is applied, the molecules start to switch so that the polarization is parallel to the applied field in most of the cell except near the top plate where the strong anchoring condition, $\phi(L, t) = \pi$, is imposed. Figure 2 depicts the director configuration with $e = 0.75$ in the left and $e = 1.3$ in the right columns, respectively. The flow effect is neglected in simulations in the first row, and it is included in the second and third rows. The control parameter A is close to 0 in the middle, while A is close to 1 in the third row. As we may expect from (4.25), the simulations in first and second rows depict almost the same switching time, while the third row describes faster switching dynamics.

In Figure 3, we numerically investigate the switching behavior when the sign of the electric field is alternating, proceeding as follows: we first obtain the director profile of the equilibrium state in a positive electric field and then impose it as an initial condition of the problem with an applied negative electric field. In [21], the authors investigated the switching time for the static model when alternating fields are applied.

The simulations show that the predicted switching process is faster when the flow is taken into consideration. We note that the switching is already faster even with a very small, but nonzero, value of A.

Zou, Clark, and Carlsson in [30] also performed numerical simulations for reorientation dynamics with various boundary conditions, based on the model proposed by Leslie, Stewart, and Nakagawa [18]. In bookshelf geometry, they showed that the switching process is generally faster when backflow is present. They also numerically confirmed that the control parameter found in [5] is a measure of the contribution of the backflow effects. The control parameter in [30] is defined as the average of $1-h(\theta)$ over θ. In the previous section, we also identified an analogous control parameter, which is dependent only on parameters of the problem, but we obtained bounds on it that rigorously allow us to quantify the backflow effects in the switching time. In particular, the upper bound on A yields an optimality condition on the parameters.

5. Conclusion. In this paper, we presented a nonlinear continuum theory of smectic C* liquid crystals. Since the smectic C liquid crystals have molecules tilted with respect to the layers, we use both the director and the layer functions as variables in the hydrodynamic theory. For the general framework, we employed the approach by Ericksen and Leslie for the hydrodynamic theory of the nematic liquid crystals. Also, motivated by the work of W. E on the continuum theory of the smectic A liquid crystals, we obtained the dynamic equations for the director $\mathbf{n}$ and the layer variable φ.

We applied the model to study the switching dynamics between two states with opposite polarization in the homeotropic geometry. Even though there are 22 viscosity coefficients in our hydrodynamic theory, the system of equations reduces to two equations with only four viscosity constants, η_1, η_2, η_3, and β_1. These constants are further constrained by the entropy inequality. We understand the molecular reorientation via the propagation of a traveling wave. We proved the existence and uniqueness of the traveling wave solution and further analyzed the speed of the front. We showed that the flow generally makes the switching faster and that there is a control parameter that determines the importance of the flow effect. This analysis was confirmed by the numerical simulations.

Acknowledgment. We would like to thank Professor Jie Shen of Purdue University for his help on numerical simulations.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[2] R. D. Benguria and M. C. Depassier, *Speed of fronts of the reaction-diffusion equation*, Phys. Rev. Let., 77 (1996), pp. 1171–1173.

[3] G. I. Blake and F. M. Leslie, *A backflow effect in smectic C liquid crystals*, Liq. Cryst., 25 (1998), pp. 319–327.

[4] S. Chandrasekhar, *Liquid Crystals*, 2nd ed., Cambridge University Press, UK, 1992.

[5] T. Carlson, N. A. Clark, and Z. Zou, *Theoretical studies of the influence of backflow on the dynamical behavior of Fredricks transition of a ferroelectric smectic C* liquid crystal in the bookshelf geometry*, Liq. Crys., 15 (1993), pp. 461–477.

[6] J.-H. Chen and T. C. Lubensky, *Landau-Ginzburg mean-field theory for the nematic to smectic-C and nematic to smectic-A phase transitions*, Phys. Rev. A, 14 (1976), pp. 1202–1207.

[7] P. E. Cladis and Van Saarloos, *Some nonlinear problems in anisotropic systems*, solicited chapter for Solitons in Liquid Crystals, L. Lam and J. Prost, eds., Springer-Verlag, New York, 1990, pp. 110–150.

[8] P. G. De Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Clarendon Press, Oxford, UK, 1993.

[9] W. E, *Nonlinear continuum theory of smectic-A liquid crystals*, Arch. Ration. Mech. Anal., 137 (1997), pp. 159–175.

[10] J. L. Ericksen, *Hydrostatic theory of liquid crystals*, Arch. Ration. Mech. Anal., 9 (1962), pp. 371–378.

[11] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equation to traveling wave solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.

[12] H. Ikeda, M. Mimura, and Y. Nishiura, *Global bifurcation phenomena of traveling wave solutions for some bistable reaction-diffusion systems*, Nonlinear Anal., 13 (1989), pp. 507–526.

[13] S. Joo and D. Phillips, *The phase transitions from chiral nematic toward smectic liquid crystals*, Comm. Math. Phys., 269 (2007), pp. 369–399.

[14] M. Kleman and O. Parodi, *Some constitutive equations for liquid crystals*, Arch. Ration. Mech. Anal., 28 (1968), pp. 265–283.

[15] B. Kazmierczak and Z. Peradzynski, *Heteroclinic solutions for a system of strongly coupled ODEs*, Math. Methods. Appl. Sci., 19 (1996), pp. 451–461.

[16] S. T. Lagerwell, *Ferroelectric and Antiferroelectric Liquid Crystals*, Wiley-VCH, Weinheim, 1999.

[17] F. Leslie, *Theory of flow phenomena in liquid crystals*, Advances in Liquid Crystals, 4 (1979), pp. 1–81.

[18] F. M. Leslie, I. W. Stewart, and M. Nakagawa, *A continuum theory for smectic C liquid crystals*, Mol. Cryst. Liq. Cryst., 198 (1991), pp. 443–454.

[19] T. C. Lubensky and S. R. Renn, *Abrikosov dislocation lattice in a model of the cholesteric to smectic-A transition*, Phys. Rev. A, 38 (1988), pp. 2132–2147.

[20] I. Luk'yanchuk, *Phase transition between the cholesteric and twist grain boundary C phases*, Phys. Rev. E, 57 (1998), pp. 574–581.

[21] J. E. Maclennan, M. A. Handschy, and N. A. Clark, *Solitary waves in ferroelectric liquid crystals*, Phys. Rev. A, 34 (1986), pp. 3554–3557.

[22] O. Parodi, *Stress tensor for a nematic liquid crystal*, J. Physique, 31 (1970), p. 581.

[23] J. Park and M. C. Calderer, *Variational problems and modeling of ferroelectricity in chiral smectic liquid crystals*, in Modeling of Soft Matter, IMA Vol. Math. Appl. 141, Springer-Verlag, New York, 2005, pp. 169–188.

[24] J. Park and M. C. Calderer, *Analysis of nonlocal electrostatic effects in chiral smectic C liquid crystals*, SIAM J. Appl. Math., 66 (2006), pp. 2107–2126.

[25] W. van Saarloos, M. van Hecke, and R. Holyst, *Front propagation into unstable and metastable states in smectic C* liquid crystals: Linear and nonlinear marginal stability analysis*, Phys. Rev. E, 52 (1995), p. 1773.

[26] P. Schiller, G. Pelzl, and D. Demus, *Bistability and domain wall motion in smectic C phases induced by strong electric fields*, Liquid Crystals, 2 (1987), pp. 21–30.

[27] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers for the second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[28] J. Shen and S. Wang, *A fast and accurate numerical scheme for the primitive equations of the atmosphere*, SIAM J. Numer. Anal., 36 (1999), pp. 719–737.

[29] I. W. Stewart, *The Static and Dynamic Continuum Theory of Liquid Crystals*, Taylor & Francis, London, 2004.

[30] Z. Zou, N. A. Clark, and T. Carlsson, *Influence of backflow on the reorientation dynamics of ferroelectric liquid crystals*, Phys. Rev. E, 49 (1994), p. 3021.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 810–829

OPTIMAL SELLING RULES IN A REGIME-SWITCHING EXPONENTIAL GAUSSIAN DIFFUSION MODEL*

P. ELOE†, R. H. LIU†, M. YATSUKI†, G. YIN‡, and Q. ZHANG§

Abstract. This paper develops optimal selling rules in asset trading using a regime-switching exponential Gaussian diffusion model. The optimization problem is solved by a combined approach of boundary value problems and probabilistic analysis. A system of linear differential equations with variable coefficients and two-point boundary conditions, satisfied by the objective function of the problem, is derived. The existence and uniqueness of the solution are proved. A closed-form solution in terms of Weber functions is obtained for one-dimensional cases. For m-dimensional cases, a stochastic recursive algorithm for numerically searching the optimal value is developed. Numerical results are reported.

Key words. optimal selling rule, Markov chain, regime-switching, Gaussian diffusion, boundary value problem, stochastic recursive algorithm

AMS subject classifications. 91B26, 91B28, 60J27, 62L20

DOI. 10.1137/060652671

1. Introduction.

This paper develops an optimal selling rule in asset trading using a regime-switching exponential Gaussian diffusion model for asset price. A selling rule is specified by two threshold levels—an upper level (greater than the purchase price) for the profit target and a lower level (less than the purchase price) for the stop-loss limit. The asset is sold once its price hits either level. Our objective in this study is to obtain a pair of optimal threshold levels that maximize a prespecified objective function which reflects the investment goal and/or risk attitude of investors.

Recently, considerable attention has been drawn to regime-switching models in financial mathematics which aim to include the influence of macroeconomic factors on the individual asset price behavior. In this setting, asset prices are dictated by a number of stochastic differential equations coupled by a finite-state Markov chain, which represents various randomly changing economical factors. Model parameters (drift and volatility coefficients) are assumed to depend on the Markov chain. Regime-switching models have been used in derivative pricing (see Buffington and Elliott [2], Guo [12], Guo and Zhang [13], and Yao, Zhang, and Zhou [22] among others), for interest rates and bond prices (see Bansal and Zhou [1] and Dai, Singleton, and Yang [6] among others), and in modeling commodity and electricity prices (see Clewlow and Strickland [3], Erlwein, Benth, and Mamon [9], Kluge [16], and Lucia and Schwartz [18] among others).

Along another line, Zhang [26] studied an optimal selling rule for stock liquidation using a regime-switching geometric Brownian motion (GBM) model. In [26], a method that combines differential equation with probabilistic analysis was developed;

*Received by the editors February 21, 2006; accepted for publication (in revised form) September 3, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/siap/69-3/65267.html

†Department of Mathematics, University of Dayton, 300 College Park, Dayton, OH 45469-2316 (paul.eloe@notes.udayton.edu, ruihua.liu@notes.udayton.edu, Masako.Yatsuki@notes.udayton.edu).

‡Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu).

§Department of Mathematics, Boyd GSRC, The University of Georgia, Athens, GA 30602-7403 (qingz@math.uga.edu).

an analytical solution for the two-regime case was obtained, and optimization techniques for deterministic functions were used to find the optimal thresholds. However, when the number of regimes exceeds two, the analytical solutions are difficult to obtain, and thus the deterministic optimization approaches are not applicable anymore. To find a feasible solution, Yin, Liu, and Zhang [23] took a different approach, namely, using stochastic approximation algorithms. By focusing on threshold-type strategies, recursive algorithms using Monte Carlo simulation were developed in [23]. Convergence and the rates of convergence of the algorithms were proved. The stochastic algorithms were tested by using both simulations and market data (see Yin, Liu, and Zhang [23], and Yin et al. [25] for more details).

In this work, we extend the aforementioned optimal selling rule study for the regime-switching GBM model to a class of regime-switching exponential Gaussian diffusion models that include the GBM and regime-switching GBM models as special cases. The new mathematical model is presented first, and its connection with other models is then noted. An objective function associated with the optimization problem is defined next. Consequently, a system of linear differential equations with two boundary conditions, satisfied by the objective function, is derived. We point out a significant difference between the system considered in this paper and that of Zhang [26]. That is, the coefficients of the differential equations are no longer constant. Therefore, solutions from [26] cannot be used in this paper. We develop a different approach. The existence of a solution to the variable coefficient boundary value problem is proved by adopting a method of upper and lower solutions that use the Green's function of the associated homogeneous system. The uniqueness of the solution is established by applying Dynkin's formula. In addition, a numerical method to construct a sequence of increasing functions (lower solution approximation) and a sequence of decreasing functions (upper solution approximation) is developed. The second part of the paper is concerned with stochastic optimization methods. We develop a recursive algorithm which provides a feasible solution for searching the best selling rules and is particularly applicable to models with large state spaces.

The rest of the paper is organized as follows. Section 2 presents the regime-switching model and the precise formulation of the selling rule problem. The differential equations and boundary values satisfied by the objective function of the optimization problem is derived. Section 3 establishes the existence and uniqueness of the solution to the problem. Section 4 is concerned with stochastic recursive algorithms. The selling rule problem is reformulated as a stochastic optimization problem. A recursive algorithm for searching the optimal thresholds using gradient estimation and projection procedure is developed. Conditions for convergence of the algorithm are provided. Numerical results are reported. Finally, the paper is concluded with further remarks in section 5.

2. Problem formulation. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the underlying probability space, upon which all stochastic processes are defined. Let $\alpha(t)$ be a continuous-time Markov chain taking values in $\mathcal{M} := \{1, \dots, m\}$, a finite state space. The states represent general market trends and other economic factors (called "state of the world" or "regime") and are labeled by integers 1 to m, where m is the total number of regimes considered for the economy. For example, with $m = 2$, $\alpha(t) = 1$ may stand for an up market and $\alpha(t) = 2$ a down market. Let $B(t)$ be a real-valued standard Brownian motion. Assume that $\alpha(t)$ is independent of $B(t)$.

Let $S(t)$ be the asset price at time $t \geq 0$,

$$S(t) = S_0 \exp(X(t)), \quad t \geq 0, \tag{2.1}$$

where $S_0 > 0$ denotes the asset price at $t = 0$ (i.e., $S(0) = S_0$), and $X(t)$ is the solution of the stochastic differential equation

$$\begin{cases} dX(t) = [b(\alpha(t)) + \mu(\alpha(t))X(t)]dt + \sigma(\alpha(t))dB(t), \\ X(0) = 0. \end{cases} \tag{2.2}$$

Note that the coefficients $b(\alpha(t))$, $\mu(\alpha(t))$, and $\sigma(\alpha(t))$ in (2.2) all depend on $\alpha(t)$, indicating that they can take different values for different regimes. We assume that $b(i) \geq 0$ and $\sigma(i) > 0$ for each $i \in \mathcal{M}$. Before introducing the optimal selling rule problem, we make three remarks regarding the model given by (2.1) and (2.2).

Remark 2.1. Consider a special case in which there is only one state for $\alpha(t)$, i.e., $m = 1$. Then $\alpha(t) = 1$ for all $t \geq 0$. In this case, these is no regime switching, and we can write $b(\alpha(t)) = b$, $\mu(\alpha(t)) = \mu$, and $\sigma(\alpha(t)) = \sigma$, where b, μ, and σ are constants. Then (2.2) becomes an Ornstein–Uhlenbeck process,

$$dX(t) = [b + \mu X(t)]dt + \sigma dB(t). \tag{2.3}$$

In particular, if we assume that $\mu < 0$ and let $\kappa = -\mu$ and $\theta = b/\kappa$, then we have the well-known Vasicek model [21] for interest rates, namely,

$$dr(t) = \kappa[\theta - r(t)]dt + \sigma dB(t),$$

where $r(t) := X(t)$ denotes the instantaneous spot rate at time $t \geq 0$, θ is the mean-reverting level, κ is the rate at which $r(t)$ is pulled back to the level θ, and σ is the volatility of $r(t)$. Also note that the solution of (2.3) is given by

$$X(t) = \frac{b}{\mu}(e^{\mu t} - 1) + \sigma \int_0^t e^{\mu(t-s)} dB(s), \tag{2.4}$$

which is a Gaussian process. Consequently, the asset price $S(t) = S_0 \exp(X(t))$ becomes an exponential Gaussian process. However, when there is more than one state for $\alpha(t)$, i.e., $m \geq 2$, then $X(t)$ will no longer be a Gaussian process. Instead, it is a mixture of m Gaussian processes. We use the term regime-switching exponential Gaussian diffusion model for (2.1) and (2.2) in this paper that generalizes the Ornstein–Uhlenbeck process.

Remark 2.2. Set $\mu(i) \equiv 0$ and let $b(i) = \nu(i) - \frac{1}{2}\sigma^2(i)$ for $i \in \mathcal{M}$ in (2.2). Then the model given by (2.1) and (2.2) is reduced to the regime-switching GBM model considered in [26] with drift $\nu(\alpha(t))$ and volatility $\sigma(\alpha(t))$, i.e.,

$$\frac{dS(t)}{S(t)} = \nu(\alpha(t))dt + \sigma(\alpha(t))dB(t),$$

which includes the commonly used log-normal model as a special case (when $m = 1$). Thus the model we consider in this paper further generalizes the (regime-switching) log-normal model. The selling rule problem for the case of $\mu(i) \equiv 0$ for $i \in \mathcal{M}$ has already been handled in Zhang [26]. In what follows we will focus on $\mu(i) \neq 0$; see also Remark 2.6.

Remark 2.3. Note that if $m = 1$, i.e., without regime switching, then the model given by (2.1) and (2.2) becomes a particular member of the class of affine diffusion models (see Duffie, Filipović, and Schachermayer [8] for definition of affine models). For $m > 1$, the model generalizes the affine model by adding a Markovian regime switching.

To continue our studies, consider a threshold-type selling rule specified by a pair of numbers z_1 and z_2 with $-\infty < z_1 \le 0 \le z_2 < \infty$. Define a stopping time τ by

$$\tau = \inf\{t > 0 : \; X(t) \notin (z_1, z_2)\}. \tag{2.5}$$

Let $S_L = S_0 e^{z_1}$ and $S_U = S_0 e^{z_2}$. Then $0 < S_L \le S_0 \le S_U < \infty$, and τ can be equivalently defined in terms of $S(t)$, i.e.,

$$\tau = \inf\{t > 0 : \; S(t) \notin (S_L, S_U)\}. \tag{2.6}$$

We call τ the selling time and S_L and S_U the lower and upper thresholds, respectively, for asset $S(t)$. That means we sell the asset at time τ either to take a profit (if S_U is reached) or to prevent further loss (if S_L is reached).

The optimal selling rule problem is to find a pair of numbers (z_1, z_2) that maximize the objective function:

$$V(z_1, z_2) = E\left\{\Phi(X(\tau)) \exp(-\rho\tau)\right\}, \tag{2.7}$$

where $\Phi(x)$ is a prespecified utility function and $\rho > 0$ is a discount factor.

Remark 2.4. Depending on the investment purpose and/or the risk attitude of an investor (risk-neutral or risk-averse), an appropriate utility function $\Phi(x)$ can be used in the objective (2.7). For instance, if we choose $\Phi(x) = e^x - 1$, then we can rewrite the objective function (2.7) as

$$V(z_1, z_2) = E\left\{\exp(-\rho\tau)\frac{S(\tau) - S_0}{S_0}\right\},$$

which gives the expectation of the discounted percentage return. By maximizing this objective function, one seeks the maximum percentage return, a common index used in evaluating investment performance.

Remark 2.5. The selling rule problem we consider in this paper is an optimal stopping problem. Note that the objective function (2.7) is determined by the first hitting time τ of process $X(t)$ at the double barriers z_1, z_2. When the log-normal model (without regime-switching) is specified for the underlying asset price, a probabilistic approach can be used to obtain the distribution function of the stopping time τ (see Karatzas and Shreve [15] and Steele [20] for extensive discussions on the probabilistic methods and results), and, consequently, an analytical objective function can be derived. However, when the new regime-switching model is used, it is difficult to obtain the distribution function of τ; thus the "pure" probabilistic approach does not work. We resort to methods of differential equations together with probabilistic approaches to solve the problem.

To proceed, we derive a two-point boundary value problem associated with (2.7). For a given real number z, consider the process $\xi(t)$ that is the solution of

$$d\xi(t) = [b(\alpha(t)) + \mu(\alpha(t))\xi(t)]dt + \sigma(\alpha(t))dB(t), \quad \xi(0) = z.$$

Then $\xi(t) = X(t)$ if $z = 0$. For each $z \in [z_1, z_2]$, define a stopping time:

$$\tau(z) = \inf\{t > 0 : \; \xi(t) \notin (z_1, z_2)\}.$$

Note we use $\tau(z)$ to indicate the z dependence of the stopping time. Let

$$v(z, i) = E\left\{\Phi(\xi(\tau(z))) \exp(-\rho\tau(z)) \middle| \alpha(0) = i, \xi(0) = z\right\}. \tag{2.8}$$

Then the objective function (2.7) can be written in terms of $v(z,i)$ as

$$V(z_1,z_2)=\sum_{i=1}^{m}p_iv(0,i), \tag{2.9}$$

where $p_i = P\{\alpha(0) = i\}$, $i = 1,\dots,m$, assumed given, is the initial probability distribution of the Markov chain $\alpha(\cdot)$.

Let matrix $Q = (q_{ij})_{m\times m}$ be the generator of the Markov chain $\alpha(\cdot)$. From Markov chain theory (see, for example, Yin and Zhang [24]), the entries q_{ij} of Q satisfy (i) $q_{ij} \geq 0$ if $j \neq i$; (ii) $\sum_{j=1}^{m} q_{ij} = 0$ for each $i = 1,\dots,m$. Moreover,

$$\lim_{\Delta t\to 0^+}\frac{P(\Delta t)-I}{\Delta t}=Q, \tag{2.10}$$

where $P(\Delta t) = (p_{ij}(\Delta t))_{m\times m} = (P\{\alpha(\Delta t) = j|\alpha(0) = i\})_{m\times m}$ is the transition probability matrix of $\alpha(\cdot)$, and I denotes the $m \times m$ identity matrix.

Consider a small interval Δt. Since $\xi(t)$ and $\alpha(t)$ are jointly Markovian, it follows that

$$v(z,i)=\sum_{j=1}^{m}E\left\{v(\xi(\Delta t),j)\exp(-\rho\Delta t)\right\}P\{\alpha(\Delta t)=j|\alpha(0)=i\}.$$

Expanding $v(\xi(\Delta t), j) \exp(-\rho\Delta t)$ at 0, using Itô's formula, sending $\Delta t \to 0$, and using the limit (2.10), we obtain the following system of differential equations associated with the value functions $v(z,i)$, $i = 1,\dots,m$:

$$\frac{\sigma^2(i)}{2}\frac{d^2v(z,i)}{dz^2}+[b(i)+\mu(i)z]\frac{dv(z,i)}{dz}-\rho v(z,i)+\sum_{j=1}^{m}q_{ij}v(z,j)=0 \tag{2.11}$$

for $z \in (z_1,z_2)$. The boundary conditions are given by

$$v(z_1,i)=\Phi(z_1),\quad v(z_2,i)=\Phi(z_2). \tag{2.12}$$

If the boundary value problem (2.11) and (2.12) has a smooth solution $v(z,i)$, $i = 1,\dots,m$, then, using Dynkin's formula (see, for example, Oksendal [19]), we can show that it must be given by (2.8), which implies the uniqueness of the solution. Therefore, it is necessary to establish the existence of a C^2 solution to (2.11) and (2.12). This is the task of the next section.

Remark 2.6. While a system of *constant* coefficient linear differential equations was obtained in Zhang [26] based on the regime-switching log-normal model for asset price, what we have here for the new model is a system of differential equations with *variable* coefficients. Therefore, methods used in [26] for constant coefficient systems are not applicable and we need a new approach for the analysis of (2.11) and (2.12). One of the major contributions of this paper (in the next section) is that we employ a new method and successfully prove the existence of a C^2 solution of the variable coefficient boundary value problem (2.11) and (2.12).

In what follows, we use f_x and f_{xx} to denote the first- and second-order derivatives of f with respect to x, respectively, where f is either a real-valued or a vector-valued function of x. Using this notation, we rewrite the system (2.11)–(2.12) in the following matrix form:

$$\begin{cases} AV_{zz}(z)+[B_1+B_0z]V_z(z)+CV(z)=FV(z) & \text{for } z\in(z_1,z_2),\\ V(z_1)=\Phi(z_1)\mathbb{1}_m,\ V(z_2)=\Phi(z_2)\mathbb{1}_m, \end{cases} \tag{2.13}$$

where $V(z) = (v(z,1),\dots,v(z,m))^T$, $1\!\!1_m = (1,\dots,1)^T$, $A = \frac{1}{2}\operatorname{diag}(\sigma^2(1),\dots,\sigma^2(m))$, $B_0 = \operatorname{diag}(\mu(1),\dots,\mu(m))$, $B_1 = \operatorname{diag}(b(1),\dots,b(m))$, $C = Q_d - \rho I = \operatorname{diag}(q_{11}-\rho, \dots, q_{mm}-\rho)$, and $F = Q_d - Q$ where $Q_d = \operatorname{diag}(q_{11},\dots,q_{mm})$.

3. Solution of the boundary value problem.

In this section, we assume that $\mu(i) > 0$ for $i \in \mathcal{M}$. The case of $\mu(i) < 0$ can be handled similarly. We first study the scalar system (the one-dimensional case) and derive an explicit solution. Then we prove the existence of a solution for multidimensional systems, using the one-dimensional result.

When $m = 1$, (2.13) reduces to a second-order scalar linear differential equation subject to two boundary conditions:

$$\left\{\begin{array}{l} \dfrac{\sigma^2}{2}V_{zz}(z) + [b+\mu z]V_z(z) - \rho V(z) = 0 \quad \text{for } z \in (z_1, z_2), \\ V(z_1) = \Phi(z_1),\ V(z_2) = \Phi(z_2), \end{array}\right. \tag{3.1}$$

where $V(z) = v(z,1)$, $\mu = \mu(1)$, $b = b(1)$, and $\sigma = \sigma(1)$. Set $x = \kappa_1 + \kappa_0 z$, where $\kappa_0 = \frac{\sqrt{2\mu}}{\sigma}$ and $\kappa_1 = \frac{b}{\sigma}\sqrt{\frac{2}{\mu}}$. Let $\widetilde{V}(x) = V(z)$. Then (3.1) is transformed into

$$\left\{\begin{array}{l} \widetilde{V}_{xx}(x) + x\widetilde{V}_x(x) - \lambda\widetilde{V}(x) = 0 \quad \text{for } x \in (\kappa_1+\kappa_0 z_1, \kappa_1+\kappa_0 z_2), \\ \widetilde{V}(\kappa_1+\kappa_0 z_1) = \Phi(z_1),\ \widetilde{V}(\kappa_1+\kappa_0 z_2) = \Phi(z_2), \end{array}\right. \tag{3.2}$$

where $\lambda := \rho/\mu$. To solve the homogeneous equation (3.2), we use the following transform:

$$\widetilde{V}(x) = \exp\left(-\frac{x^2}{4}\right) D(x).$$

Then $D(x)$ satisfies

$$D_{xx}(x) + \left[\frac{1}{2} - \frac{x^2}{4} - \bar{\lambda}\right] D(x) = 0, \tag{3.3}$$

where $\bar{\lambda} := 1 + \lambda > 0$. From the results presented in Darling and Siegert [7] and Finch [10], we have the following proposition.

Proposition 3.1. *The function $D^\nu(x)$ defined below (known as the parabolic cylinder function or the Weber function) satisfies the equation*

$$D^\nu_{xx}(x) + \left[\frac{1}{2} - \frac{x^2}{4} + \nu\right] D^\nu(x) = 0, \tag{3.4}$$

where

$$D^\nu(x) = \left\{\begin{array}{ll} \sqrt{\dfrac{2}{\pi}}\exp\left(\dfrac{x^2}{4}\right)\displaystyle\int_0^\infty t^\nu \exp\left(-\frac{t^2}{2}\right)\cos\left(xt - \frac{\pi\nu}{2}\right)dt, & \nu > -1, \\ \dfrac{1}{\Gamma(-\nu)}\exp\left(-\dfrac{x^2}{4}\right)\displaystyle\int_0^\infty t^{-\nu-1}\exp\left(-\frac{t^2}{2} - xt\right)dt, & \nu < 0, \end{array}\right. \tag{3.5}$$

and $\Gamma(\cdot)$ is the Gamma function. The two branches in (3.5) *agree for $-1 < \nu < 0$.*

Comparing (3.3) with (3.4), we see that one solution of (3.3) is given by

$$D(x) = D^{-\bar{\lambda}}(x) = D^{-(1+\lambda)}(x) = \frac{1}{\Gamma(1+\lambda)}\exp\left(-\frac{x^2}{4}\right)\int_0^\infty t^\lambda \exp\left(-\frac{t^2}{2} - xt\right)dt.$$

The second independent solution is given by

$$D(-x) = \frac{1}{\Gamma(1+\lambda)} \exp\left(-\frac{x^2}{4}\right) \int_0^\infty t^\lambda \exp\left(-\frac{t^2}{2} + xt\right) \, dt.$$

It follows that the solution to (3.2) is

$$\widetilde{V}(x) = C_1 \int_0^\infty t^\lambda \exp\left(-\frac{(t+x)^2}{2}\right) \, dt + C_2 \int_0^\infty t^\lambda \exp\left(-\frac{(t-x)^2}{2}\right) \, dt,$$

where C_1 and C_2 are constants to be determined using the given boundary conditions.

Consider the scalar boundary value problem defined below:

$$(3.6) \qquad \begin{cases} D_{xx}(x) + \left[\frac{1}{2} - \frac{x^2}{4} - (1+\gamma)\right] D(x) = 0 \quad \text{for } x \in (x_1, x_2), \\ D(x_1) = 0, \; D(x_2) = 0, \end{cases}$$

where $\gamma > 0$ is a fixed constant. Set

$$D_1(x) = \exp\left(-\frac{x^2}{4}\right) \int_0^\infty t^\gamma \exp\left(-\frac{t^2}{2} - xt\right) \, dt$$

and

$$D_2(x) = \exp\left(-\frac{x^2}{4}\right) \int_0^\infty t^\gamma \exp\left(-\frac{t^2}{2} + xt\right) \, dt.$$

Then D_1 and D_2 form a Descartes system of solutions for the homogeneous equation in (3.6), since $D_1 > 0$, $D_2 > 0$, and $W(D_1, D_2) > 0$ on $[x_1, x_2]$, where

$$W(D_1, D_2) = \det \begin{pmatrix} D_1 & D_2 \\ D_{1,x} & D_{2,x} \end{pmatrix}$$

denotes the Wronskian of D_1 and D_2. Thus, the equation in (3.6) is disconjugate on $[x_1, x_2]$ (see Coppel [5]). This result, coupled with the observation that the boundary conditions in (3.6) (i.e., $D(x_1) = 0$ and $D(x_2) = 0$) are two-point conjugate boundary conditions, implies two immediate corollaries which we shall employ below to establish the existence of a solution of (2.13) and to provide numerical approximations that converge monotonically to the appropriate C^2 solution.

Corollary 3.2. *There exists a Green's function $G(\gamma; x, s)$ for the boundary value problem (3.6) satisfying*

$$G(\gamma; x, s) < 0 \text{ for } (x, s) \in (x_1, x_2) \times (x_1, x_2).$$

Moreover, $G_x(\gamma; x_1, s) < 0$ for $s \in (x_1, x_2)$ and $G_x(\gamma; x_2, s) > 0$ for $s \in (x_1, x_2)$, where G_x denotes the partial derivative of G with respect to x.

Note that the Green's function G plays the role that

$$D(x) = \int_{x_1}^{x_2} G(x, s) f(s) ds$$

is the unique solution of

$$\begin{cases} D_{xx}(x) + \left[\frac{1}{2} - \frac{x^2}{4} - (1+\gamma)\right] D(x) = f(x) \quad \text{for } x \in (x_1, x_2), \\ D(x_1) = 0, \; D(x_2) = 0, \end{cases}$$

where $f(x)$ is a continuous function on $[x_1, x_2]$.

COROLLARY 3.3. *The solution of the boundary value problem*

$$\begin{cases} D_{xx}(x) + \left[\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\gamma)\right] D(x) = 0 \quad \textit{for } x \in (x_1, x_2), \\ D(x_1) > 0,\ D(x_2) > 0 \end{cases}$$

is positive on $[x_1, x_2]$.

Proof. The disconjugacy of $D_{xx}(x) + [\frac{1}{2} - \frac{x^2}{4} - (1+\gamma)]\, D(x) = 0$ on $[x_1, x_2]$ means that any nontrivial solution has at most one root (counting multiplicities) on $[x_1, x_2]$. Since the solution is strictly positive at each boundary, the desired result follows. □

Having done with the one-dimensional case, now we address the existence of a C^2 solution to the m-dimensional ($m > 1$) boundary value system (2.13). To carry out the analysis, the following assumption is needed.

ASSUMPTION 3.4.

$$\frac{\mu(1)}{\sigma^2(1)} = \frac{\mu(2)}{\sigma^2(2)} = \cdots = \frac{\mu(m)}{\sigma^2(m)}$$

and

$$\frac{b^2(1)}{\sigma^2(1)\mu(1)} = \frac{b^2(2)}{\sigma^2(2)\mu(2)} = \cdots = \frac{b^2(m)}{\sigma^2(m)\mu(m)} .$$

THEOREM 3.5. *Under Assumption* 3.4*, there exists a unique* C^2 *solution to the boundary value problem* (2.13)*.*

Proof. We employ the method of upper and lower solutions to obtain existence. Let $x = \kappa_1 + \kappa_0 z$, where $\kappa_0 = \frac{\sqrt{2\mu(i)}}{\sigma(i)}$ and $\kappa_1 = \frac{b(i)}{\sigma(i)}\sqrt{\frac{2}{\mu(i)}}$ are two constants due to Assumption 3.4. For notational brevity, in what follows, we introduce

$$\bar{\kappa}_1 = \kappa_1 + \kappa_0 z_1\ ,\ \bar{\kappa}_2 = \kappa_1 + \kappa_0 z_2. \tag{3.7}$$

Let $\widetilde{V}(x) = V(z)$. Then (2.13) is converted to the following problem:

$$\begin{cases} \widetilde{V}_{xx}(x) + x\widetilde{V}_x(x) - \widetilde{C}\widetilde{V}(x) = \widetilde{F}\widetilde{V}(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ \widetilde{V}(\bar{\kappa}_1) = \Phi(z_1)\mathbb{1}_m,\ V(\bar{\kappa}_2) = \Phi(z_2)\mathbb{1}_m, \end{cases} \tag{3.8}$$

where

$$\widetilde{C} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m),\ \ \lambda_i = \frac{\rho - q_{ii}}{\mu(i)},\ \ i = 1, \ldots, m, \tag{3.9}$$

and

$$\widetilde{F} = \begin{pmatrix} 0 & -q_{12}/\mu(1) & \cdots & -q_{1m}/\mu(1) \\ -q_{21}/\mu(2) & 0 & \cdots & -q_{2m}/\mu(2) \\ \vdots & \vdots & \cdots & \vdots \\ -q_{m1}/\mu(m) & -q_{m2}/\mu(m) & \cdots & 0 \end{pmatrix}. \tag{3.10}$$

Note that $\rho > 0$, $\mu(i) > 0$, and $q_{ii} \le 0$. Hence $\lambda_i > 0$ for $i = 1, \ldots, m$.

We use the (vector) transform $\widetilde{V}(x) = \exp(-\frac{x^2}{4})D(x)$, where

$$D(x) = (D_1(x), \dots, D_m(x))^T.$$

Then (3.8) is transformed into

$$\tag{3.11} \begin{cases} D_{xx}(x) + \bar{C}D(x) = \widetilde{F}D(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ D(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1)\mathbb{1}_m, \;\; D(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\mathbb{1}_m, \end{cases}$$

where

$$\tag{3.12} \bar{C} = \operatorname{diag}\left(\left[\frac{1}{2} - \frac{x^2}{4} - (1+\lambda_1)\right], \dots, \left[\frac{1}{2} - \frac{x^2}{4} - (1+\lambda_m)\right]\right).$$

Note that the left-hand side of the vector equation (3.11) is decoupled and, hence, diagonal. For each $i = 1, \dots, m$, the corresponding homogeneous scalar boundary value problem is given by

$$\tag{3.13} \begin{cases} D_{i,xx}(x) + \left[\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_i)\right] D_i(x) = 0 \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ D_i(\bar{\kappa}_1) = 0, \; D_i(\bar{\kappa}_2) = 0. \end{cases}$$

Let $G(\lambda_i; x, s)$ be the associated Green's function as given by Corollary 3.2. Define

$$G(x,s) = \operatorname{diag}\Big(G(\lambda_1; x, s), \dots, G(\lambda_m; x, s)\Big).$$

Then $G(x,s)$ is a Green's function of the system (3.11).

Next, define a Banach space C_m by

$$C_m[\bar{\kappa}_1, \bar{\kappa}_2] = \Big\{U = (u_1, \dots, u_m)^T : [\bar{\kappa}_1, \bar{\kappa}_2] \to \mathbb{R}^m, u_i \in C[\bar{\kappa}_1, \bar{\kappa}_2], i = 1, \dots, m\Big\}$$

with norm $\|U\| = \max_{1 \le i \le m}\{\|u_i\|_0\}$, where $\|\cdot\|_0$ denotes the usual supremum norm. Consider the partial order on $\mathbb{R}^m$:

$$V \le U \iff v_i \le u_i, \quad i = 1, \dots, m, \text{ where } U, V \in \mathbb{R}^m.$$

Using this partial order, we define a partial order on $C_m[\bar{\kappa}_1, \bar{\kappa}_2]$:

$$V \le U \iff V(x) \le U(x), x \in [\bar{\kappa}_1, \bar{\kappa}_2], \text{ where } U, V \in C_m.$$

Let $D_\Phi \in C_m$ denote the solution of the following homogeneous equation with nonhomogeneous boundary conditions:

$$\tag{3.14} \begin{cases} D_{xx}(x) + \bar{C}D(x) = 0 \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ D(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1)\mathbb{1}_m, \;\; D(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\mathbb{1}_m. \end{cases}$$

The existence of D_Φ is ensured by Corollary 3.3. Define an operator $\mathbf{K}$ on C_m by

$$\tag{3.15} (\mathbf{K}D)(x) = D_\Phi(x) + \int_{\bar{\kappa}_1}^{\bar{\kappa}_2} G(x,s)\widetilde{F}D(s)\,ds,$$

where $\widetilde{F}$ is given by (3.10).

Remark 3.6. Let $\mathbf{K}$ be defined by (3.15). Then $\mathbf{K} : C_m[\bar{\kappa}_1, \bar{\kappa}_2] \longrightarrow C^2_m[\bar{\kappa}_1, \bar{\kappa}_2]$.

The remark follows by standard properties of the diagonal structure of the Green's matrix $G(x, s)$ (see Coddington and Levinson [4, p. 192]). In fact, each scalar-valued function $G(\lambda_i; x, s)$ is continuous on triangles, $x < s$, $s < x$, and satisfies the differential equation

$$D_{i,xx}(x) + \left[\frac{1}{2} - \frac{x^2}{4} - (1 + \lambda_i)\right] D_i(x) = 0$$

on triangles, $x < s$, $s < x$, and

$$\lim_{x \to s^+} G_x(x, s) - \lim_{x \to s^-} G_x(x, s) = 1.$$

If $D \in C_m[\bar{\kappa}_1, \bar{\kappa}_2]$, then it is standard to show that $\mathbf{K}D \in C^2_m[\bar{\kappa}_1, \bar{\kappa}_2]$.

The following remark is also immediate from Corollary 3.2 and (3.15) (see Coddington and Levinson [4, p. 192] and Jackson [14, p. 99]).

Remark 3.7. $D \in C^2_m$ is a solution of the boundary value problem (3.11) if and only if $D \in C_m$ and $\mathbf{K}D = D$.

In view of Corollary 3.2 and (3.10), we have $G(x, s)\widetilde{F} \geq 0$ elementwise. Therefore, $\mathbf{K}$ is a monotonic operator; that is,

$$V \leq U \Longrightarrow \mathbf{K}V \leq \mathbf{K}U, \quad U, V \in C_m.$$

We establish upper and lower solutions of the boundary value problem (3.11), respectively. That is, (see Jackson [14]), we seek $U_0 \in C^2_m$ and $V_0 \in C^2_m$ satisfying

$$V_0 \leq U_0, \quad V_0 \leq \mathbf{K}V_0, \quad \mathbf{K}U_0 \leq U_0, \tag{3.16}$$

and

$$V_0(\bar{\kappa}_i) \leq \exp\left(\frac{(\bar{\kappa}_i)^2}{4}\right) \Phi(z_i) \mathbb{1}_m \leq U_0(\bar{\kappa}_i), \quad i = 1, 2. \tag{3.17}$$

Once we obtain the upper and lower solutions, the proof for existence of a solution is complete. To see this, define a closed and convex region $\Omega \subset C_m$ by

$$D \in \Omega \Longleftrightarrow V_0(x) \leq D(x) \leq U_0(x), \quad \bar{\kappa}_1 \leq x \leq \bar{\kappa}_2 \ .$$

The inequalities (3.16) and (3.17), coupled with the fact that $\mathbf{K}$ is monotone, imply that $\mathbf{K} : \Omega \to \Omega$. Thus, the existence of a solution D, satisfying

$$V_0(x) \leq D(x) \leq U_0(x), \qquad \bar{\kappa}_1 \leq x \leq \bar{\kappa}_2, \tag{3.18}$$

follows as an application of the Schauder fixed point theorem (Jackson [14, p. 102]).

It can be shown, using the definition (3.15) and Corollary 3.2, that V_0 is a lower solution if

$$\begin{cases} V_{0,xx}(x) + \bar{C}V_0(x) \geq \widetilde{F}V_0(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ V_0(\bar{\kappa}_1) \leq \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right) \Phi(z_1)\mathbb{1}_m, \;\; V_0(\bar{\kappa}_2) \leq \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right) \Phi(z_2)\mathbb{1}_m, \end{cases} \tag{3.19}$$

where $\bar{C}$ is the diagonal matrix defined in (3.12). Similarly, U_0 is a upper solution if the above inequalities are reversed, i.e.,

$$\begin{cases} U_{0,xx}(x) + \bar{C}U_0(x) \le \widetilde{F}U_0(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ U_0(\bar{\kappa}_1) \ge \exp\left(\frac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1)\mathbb{1}_m, \;\; U_0(\bar{\kappa}_2) \ge \exp\left(\frac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\mathbb{1}_m. \end{cases} \tag{3.20}$$

Consider the solution D_Φ of (3.14). Since D_Φ satisfies the homogeneous equation (3.14), $0 \ge \widetilde{F}D_\Phi$, and D_Φ satisfies the boundary conditions, it is readily seen that D_Φ satisfies (3.19). Thus we can choose $V_0 = D_\Phi$.

On the other hand, the upper solution can be chosen as $U_0 = (K, \dots, K)^T \in \mathbb{R}^m$, where K is a constant satisfying

$$K \ge \max\left\{\exp\left(\frac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1), \exp\left(\frac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\right\}. \tag{3.21}$$

To show that the so-chosen U_0 is indeed an upper solution, we need only to verify the first inequality in (3.20). In fact, in view of (3.9) and (3.10), substituting the constant vector $(K, \dots, K)^T$ into the inequality yields, for $i = 1, \dots, m$,

$$\frac{1}{2} - \frac{x^2}{4} - (1 + \lambda_i) = \frac{1}{2} - \frac{x^2}{4} - \left(1 + \frac{\rho}{\mu(i)} - \frac{q_{ii}}{\mu(i)}\right) \le -\sum_{j \ne i} \frac{q_{ij}}{\mu(i)},$$

which is true in view of $\sum_{j=1}^m q_{ij} = 0$ and $\mu(i) > 0$. This completes the proof of the existence of a C^2 solution. □

Remark 3.8. Define $V_{k+1} = \mathbf{K}V_k$, $U_{k+1} = \mathbf{K}U_k$, $k = 0, 1, 2 \dots$. Then it follows that

$$V_k \le V_{k+1} \le U_{k+1} \le U_k, \quad k \ge 0.$$

This string of inequalities is immediate from the monotonicity of $\mathbf{K}$. Consequently, there exist functions $\bar{V}, \bar{U}$ such that $\{V_k\} \uparrow \bar{V}$, $\{U_k\} \downarrow \bar{U}$ (pointwise and componentwise) as $k \to \infty$. Moreover, by Dini's theorem, the convergence is uniform in x. So $\bar{V}, \bar{U} \in C_m$. Applying operator (3.15) to V_k (resp., U_k) and letting $k \to \infty$, we have $\mathbf{K}\bar{V} = \bar{V}$ and $\mathbf{K}\bar{U} = \bar{U}$. Therefore, both $\bar{V}$ and $\bar{U}$ are the solutions of (3.11). From Remark 3.6, we know both $\bar{V}$ and $\bar{U}$ are C_m^2 functions. The uniqueness of the solution implies that $\bar{V} = \bar{U}$.

Remark 3.9. From the proof of Theorem 3.5, we also see that the C^2 solution of the system (2.13) (and therefore the objective function (2.9)) is continuous with respect to the boundary points z_1 and z_2.

Now we study the optimality of the objective function (2.9). We make the following assumption on z_1 and z_2; see Zhang [26] for further discussions.

Assumption 3.10.

$$a_1 \le z_1 \le b_1, \quad a_2 \le z_2 \le b_2,$$

where a_1, b_1, a_2, b_2 are prespecified constants satisfying $-\infty < a_1 < b_1 < 0 < a_2 < b_2 < \infty$.

Theorem 3.11. *Under Assumptions* 3.4 *and* 3.10*, the following assertions hold:*

1. *For each $1 \le i \le m$, $v(z, i) \in C^2$ and is the unique solution to* (2.11) *and* (2.12).

2. *For each fixed pair* (z, i)*,* $v(z, i)$ *is a continuous function of* (z_1, z_2) *on* $[a_1, b_1] \times [a_2, b_2]$*.*
3. *There exists an optimal pair* $(z_1^*, z_2^*) \in [a_1, b_1] \times [a_2, b_2]$ *that maximizes the objective function* (2.9)*.*

Proof. Parts 1 and 2 are obtained by Theorem 3.5 together with Dynkin's formula. Part 3 follows from the compactness of $[a_1, b_1] \times [a_2, b_2]$. □

We provide a numerical example to demonstrate the approximation process proposed in Remark 3.8.

Example 3.12. Consider a two-dimensional system ($m = 2$) and construct the two sequences of approximation solutions (upper and lower) by iteratively solving the corresponding boundary value problems. We numerically solve these equations and graphically display the convergence of the two sequences.

When $m = 2$, the system (3.11) can be written componentwise as

$$\tag{3.22} \begin{cases} D_{1,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_1)\right) D_1(x) = -\dfrac{q_{12}}{\mu(1)} D_2(x), \\ D_{2,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_2)\right) D_2(x) = -\dfrac{q_{21}}{\mu(2)} D_1(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ D_i(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right) \Phi(z_1), \quad D_i(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right) \Phi(z_2), \quad i = 1, 2. \end{cases}$$

We first find the solution $D_\Phi = (D_{1,\Phi}, D_{2,\Phi})^T$ of the associated homogeneous equations with nonhomogeneous boundary conditions, i.e.,

$$\tag{3.23} \begin{cases} D_{1,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_1)\right) D_1(x) = 0, \\ D_{2,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_2)\right) D_2(x) = 0 \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ D_i(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right) \Phi(z_1), \quad D_i(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right) \Phi(z_2), \quad i = 1, 2. \end{cases}$$

To make the expression compact, let

$$W_\lambda(x) = \exp\left(-\frac{x^2}{4}\right) \int_0^\infty t^\lambda \exp\left(-\frac{t^2}{2} - xt\right) dt.$$

Then we have

$$\tag{3.24} D_{1,\Phi}(x) = C_1 W_{\lambda_1}(x) + C_2 W_{\lambda_1}(-x),$$

where the two constants C_1, C_2 are determined by the given pair of boundary conditions,

$$C_1 = \frac{W_{\lambda_1}(-\bar{\kappa}_2) \exp\left(\frac{(\bar{\kappa}_1)^2}{4}\right) \Phi(z_1) - W_{\lambda_1}(-\bar{\kappa}_1) \exp\left(\frac{(\bar{\kappa}_2)^2}{4}\right) \Phi(z_2)}{W_{\lambda_1}(\bar{\kappa}_1) W_{\lambda_1}(-\bar{\kappa}_2) - W_{\lambda_1}(-\bar{\kappa}_1) W_{\lambda_1}(\bar{\kappa}_2)},$$

$$C_2 = \frac{W_{\lambda_1}(\bar{\kappa}_1) \exp\left(\frac{(\bar{\kappa}_2)^2}{4}\right) \Phi(z_2) - W_{\lambda_1}(\bar{\kappa}_2) \exp\left(\frac{(\bar{\kappa}_1)^2}{4}\right) \Phi(z_1)}{W_{\lambda_1}(\bar{\kappa}_1) W_{\lambda_1}(-\bar{\kappa}_2) - W_{\lambda_1}(-\bar{\kappa}_1) W_{\lambda_1}(\bar{\kappa}_2)}\ .$$

Replacing λ_1 in the equations for $D_{1,\Phi}$ with λ_2 yields $D_{2,\Phi}$. Thus we obtain an analytical lower solution $V_0 = D_\Phi$.

Starting at $V_0 = (V_{0,1}, V_{0,2})^T$, the approximate sequence $V_k = (V_{k,1}, V_{k,2})^T$, $k \geq 1$, can be constructed by iteratively solving the following two-point boundary value problem:

$$\tag{3.25} \begin{cases} V_{k+1,1,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_1)\right) V_{k+1,1}(x) = -\dfrac{q_{12}}{\mu(1)} V_{k,2}(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ V_{k+1,1}(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1), \quad V_{k+1,1}(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2), \end{cases}$$

$$\tag{3.26} \begin{cases} V_{k+1,2,xx}(x) + \left(\dfrac{1}{2} - \dfrac{x^2}{4} - (1+\lambda_2)\right) V_{k+1,2}(x) = -\dfrac{q_{21}}{\mu(2)} V_{k,1}(x) \quad \text{for } x \in (\bar{\kappa}_1, \bar{\kappa}_2), \\ V_{k+1,2}(\bar{\kappa}_1) = \exp\left(\dfrac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1), \quad V_{k+1,2}(\bar{\kappa}_2) = \exp\left(\dfrac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\ . \end{cases}$$

The same process, starting at the upper solution $U_0 = (K, K)^T$, will produce the other sequence U_k, $k \geq 0$. In view of (3.21), we choose

$$K = \max\left\{\exp\left(\frac{(\bar{\kappa}_1)^2}{4}\right)\Phi(z_1),\ \exp\left(\frac{(\bar{\kappa}_2)^2}{4}\right)\Phi(z_2)\right\}.$$

We used a box method (see Zwillinger [27]) to solve (3.25) and (3.26). Various parameters for the numerical experiment were chosen as follows:

$$\mu(1) = 0.1, \quad \mu(2) = 0.2, \quad b(1) = 0, \quad b(2) = 0, \quad \sigma^2(1) = 0.25, \quad \sigma^2(2) = 0.5,$$

$$Q = (q_{ij}) = \begin{pmatrix} -2 & 2 \\ 3 & -3 \end{pmatrix}, \quad \rho = 1, \quad z_1 = -1, \quad z_2 = 1, \quad \Phi(x) \equiv 1.$$

Figure 1 displays a number of upper and lower approximation solutions. It demonstrates that the upper and lower approximate sequences converge to a common solution, which is the unique solution to the boundary value system.

4. Stochastic optimization method. Except for the special one-dimensional case, it is very difficult to obtain the analytical representation of the objective function (2.7). Thus finding a systematic way of obtaining the optimal threshold values becomes an important task. To search for the optimal thresholds, we develop stochastic recursive approximation algorithms in this section. To this end, we reformulate the task of finding optimal thresholds as a stochastic approximation or stochastic optimization problem. For a general approach to stochastic approximation methods, the reader is referred to Kushner and Yin [17] for an up-to-date account of stochastic approximation.

4.1. Optimization problem and stochastic approximation algorithms. In lieu of using the differential equation method, we convert the optimal stopping problem to a stochastic optimization problem. The rationale is based on using a threshold-type strategy, and the underlying problem can be stated as

$$\tag{4.1} \text{Problem } \mathcal{P}: \begin{cases} \text{Find argmax } \varphi(z) = E\{\Phi(X(\tau))\exp(-\rho\tau)\}, \\ z = (z_1, z_2)^T \in [a_1, b_1] \times [a_2, b_2], \end{cases}$$

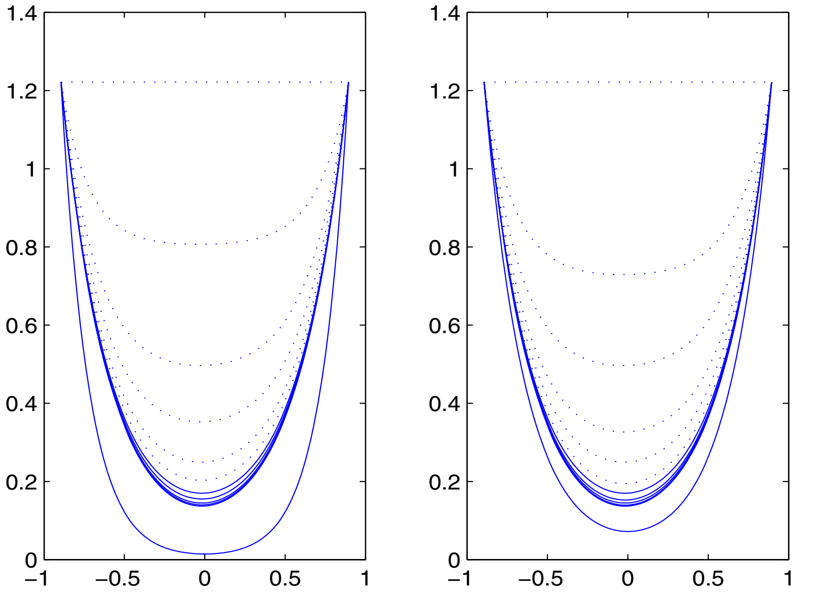


FIG. 1. *Approximation sequences and convergence. The dotted lines are the upper approximation sequences, and the solid lines are the lower approximation sequences. The left graph is for $D_1(x)$, and the right graph is for $D_2(x)$.*

where we use $\varphi(z)$ for the objective function $V(z_1, z_2)$ defined in (2.7), and τ is the stopping time defined by (2.5). Our objective is to find the optimal vector-valued threshold value for the constraint optimization problem $\mathcal{P}$.

To approximate the optimal threshold value $z^* = (z_1^*, z_2^*)^T$, we construct a recursive algorithm

$$z_{n+1} = z_n + \{\text{step size}\} \cdot \{\text{gradient estimate of } \varphi(z)\}, \tag{4.2}$$

where $z_n = (z_{n,1}, z_{n,2})^T$ denote the threshold values at the nth iteration. The step size is typically a decreasing sequence of real numbers satisfying certain conditions.

To implement (4.2), we need to construct gradient estimates of the objective function $\varphi(z)$ either by observing the real data with noisy measurements or by using a simulation. We use ξ to denote the collective random factors (including the Brownian motion, the Markov chain, and other observation noise or simulation of random effects from random seeds) so that each realization of ξ uniquely determines a sample path of the asset price dynamics (2.2) as well as the stopping time τ (2.5) for a fixed value of z. At the nth iteration, suppose the threshold values are $z_n = (z_{n,1}, z_{n,2})^T$. Let $\widetilde{\varphi}(z_n, \xi_n)$ denote the value of the discounted utility function either observed or simulated using the sample path associated with ξ_n. We assume that $E\{\widetilde{\varphi}(z, \xi_n)\} = \varphi(z)$.

Let $\Delta\widetilde{\varphi}(z_n, \xi_n) = (\Delta_1\widetilde{\varphi}(z_n, \xi_n), \Delta_2\widetilde{\varphi}(z_n, \xi_n))^T$ denote the sample path gradient estimates using a finite difference approximation, where, for $i = 1, 2$,

$$\Delta_i\widetilde{\varphi}(z_n, \xi_n) = \frac{\widetilde{\varphi}(z_n + \delta_n e_i, \xi_n) - \widetilde{\varphi}(z_n - \delta_n e_i, \xi_n)}{2\delta_n}, \tag{4.3}$$

$e_1 = (1, 0)^T$ and $e_2 = (0, 1)^T$ are the standard unit vectors, and $\{\delta_n\}$ is a sequence of positive real numbers tending to 0 and satisfying certain conditions.

Remark 4.1. The following points are worth noting.

(1) In (4.3), we use the same sample path generated by ξ_n for calculations of the function $\widetilde{\varphi}$ at different z values. This is because when Monte Carlo simulation is used to calculate a finite difference gradient approximation, using the same random numbers in calculating the two function values can reduce the variance of the estimator (see, for example, Glasserman [11]). The common random number generators can be effectively used in conjunction with stochastic approximation methods; see Kushner and Yin [17, pp. 15, 143].

(2) In the above construction of gradient estimates, instead of one simulation run, we could use multiple replications. We could use (2.2) to generate n_0 independent sample paths of $X(t)$. For each sample path, we find the value of τ, i.e., the first exit time of $X(t)$ from the interval $(z_{n,1}, z_{n,2})$. Then we construct the gradient estimates using n_0 different random seeds and then average them out. In lieu of one replication, we then use the average of n_0 replications as the gradient estimator. The advantage is that the result will be smoother. However, if we deal with real data, this idea cannot be implemented. For simplicity, we do not write the expression but refer the reader to [23] for further details.

The stochastic recursive algorithm (4.2) takes the form

$$z_{n+1} = z_n + \varepsilon_n \Delta\widetilde{\varphi}(z_n, \xi_n), \tag{4.4}$$

where $\{\varepsilon_n\}$ is a sequence of real numbers known as step sizes satisfying $0 \le \varepsilon_n \to 0$ and $\varepsilon_n/\delta_n \to 0$ as $n \to \infty$, and $\sum_n \varepsilon_n = \infty$. To ensure the boundedness of the iterates, similarly to Yin, Liu, and Zhang [23] (see also [17, p. 121]), we use the following modified stochastic approximation algorithm for the constrained problem $\mathcal{P}$:

$$z_{n+1} = \Pi[z_n + \varepsilon_n \Delta\widetilde{\varphi}(z_n, \xi_n)], \tag{4.5}$$

or, in a component form,

$$z_{n+1,i} = \Pi_{[a_i,b_i]}[z_{n,i} + \varepsilon_n \Delta_i \widetilde{\varphi}(z_n, \xi_n)] \quad \text{for} \quad i = 1, 2,$$

where the projection Π is defined as, for each real value x,

$$\Pi_{[a_i,b_i]}(x) = \begin{cases} a_i & \text{if } x < a_i, \\ b_i & \text{if } x > b_i, \\ x & \text{otherwise.} \end{cases}$$

The idea is as follows: For each component i, after the update $z_{n,i} + \varepsilon_n \Delta_i \widetilde{\varphi}(z_n, \xi_n)$ is obtained, we compare this value with the bounds a_i and b_i. If the updated value is smaller than the lower value a_i, reset the value to a_i; if it is greater than the upper value b_i, reset it to b_i; otherwise keep the value as it was. Note that in view of the techniques in [17, Chapter 5], the projection algorithm may be rewritten as

$$z_{n+1} = z_n + \varepsilon_n \Delta\widetilde{\varphi}(z_n, \xi_n) + \varepsilon_n R_n, \tag{4.6}$$

where $\varepsilon_n R_n = z_{n+1} - z_n - \varepsilon_n \Delta\widetilde{\varphi}(z_n, \xi_n)$, known as reflection term, is the minimal force needed to bring the iterates back to the constrained region if they ever escape from there.

In what follows, we present sufficient conditions guaranteeing the convergence of the algorithm. For analysis purposes only, define

$$\begin{aligned} &\psi_n = \Delta\widetilde{\varphi}(z_n, \xi_n) - E_n \Delta\widetilde{\varphi}(z_n, \xi_n), \\ &\zeta_{n,i} = E_n \Delta_i \widetilde{\varphi}(z_n, \xi_n) - [\varphi(z_n + \delta_n e_i) - \varphi(z_n - \delta_n e_i)], \quad i = 1, 2, \\ &b_{n,i} = \frac{\varphi(z_n + \delta_n e_i) - \varphi(z_n - \delta_n e_i)}{2\delta_n} - \frac{\partial \varphi(z_n)}{\partial z^i}, \quad i = 1, 2, \end{aligned} \tag{4.7}$$

where E_n denotes the conditional expectation with respect to $\mathcal{F}_n$, the σ-algebra generated by $\{z_0, \xi_j : j < n\}$, and $\varphi_z(z) = ((\partial/\partial z^1)\varphi(z), (\partial/\partial z^2)\varphi(z))^T$ denotes the gradient of $\varphi(\cdot)$. Above, $\zeta_{n,i}$ and $b_{n,i}$ for $i = 1, 2$ represent the noise and bias, and $\{\psi_n\}$ is a martingale difference sequence. This separation together with the expanded form of the recursion is for analysis purposes. As far as computation is concerned, only (4.5) is needed.

Write $\zeta_n = (\zeta_{n,1}, \zeta_{n,2})^T$ and $\beta_n = (b_{n,1}, b_{n,2})^T$ and note that $\zeta_n = \zeta_n(z_n, \xi_n)$. With the noise $\zeta_n(z_n, \xi_n)$ and the bias β_n defined above, algorithm (4.5) becomes

$$z_{n+1} = z_n + \varepsilon_n \varphi_z(z_n) + \varepsilon_n \frac{\psi_n}{2\delta_n} + \varepsilon_n \beta_n + \varepsilon_n \frac{\zeta_n}{2\delta_n} + \varepsilon_n R_n. \tag{4.8}$$

Denote $t_n = \sum_{i=1}^{n-1} \varepsilon_i$ and

$$m(t) = \begin{cases} n : t_n \le t < t_{n+1}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

To study the convergence of the algorithm, define a piecewise constant interpolation by $z^0(t) = z_n$ for $t \in [t_n, t_{n+1})$ and $z^n(t) = z^0(t + t_n)$ for $n > 0$. Similarly, define the interpolation for R_n. Let $\{\Delta_n\}$ be a sequence of positive real numbers tending to 0 as $n \to \infty$ such that $\sup_{j \ge n} \varepsilon_j / \Delta_n \to 0$. Select an increasing sequence $n = m_1 < m_2 < \cdots$ such that $\sum_{k=m_l}^{m_{l+1}-1} \varepsilon_k / \Delta_n \to 1$ as $n \to \infty$ uniformly in l. Then we have the following convergence result.

PROPOSITION 4.2. *Assume that $\varphi_{zz}(\cdot)$, the second partial derivative of $\varphi(\cdot)$, is continuous, that $\sup_n E|\widetilde{\varphi}(z, \xi_n)|^2 < \infty$ for each z, that the projected ordinary differential equation*

$$\dot{z}(t) = \varphi_z(z(t)) + r(t), \;\; r(t) \in C(z(t)) \tag{4.9}$$

has a unique solution for each initial condition, and that there is a unique stationary point z_* of (4.9) in $(a_1, b_1) \times (a_2, b_2)$ that is globally asymptotically stable in the sense of Liapunov. In addition, for each z in the constraint set, $\{\zeta_n(z, \xi)\}$ is uniformly integrable, and $\sum_{k=m_l}^{m_{l+1}-1} \varepsilon_k E_{m_l} \zeta(z, \xi_k)/\delta_k \to 0$ in probability. Then $z^n(\cdot)$ converges to $z(\cdot)$, the solution of the projected ordinary differential equation (4.9). Assume that $\{s_n\}$ is a sequence of real numbers satisfying $s_n \to \infty$ as $n \to \infty$. Then $z^n(s_n + \cdot)$ converges to z_* with probability 1.

In Proposition 4.2, $r(t)$ satisfies $R(t) = \int_0^t r(s) ds$, with $R(t)$ being the limit of the interpolation sequence of the projection term R_n. The set $C(z)$ is defined as follows: If z is inside $(a_1, b_1) \times (a_2, b_2)$, then $C(z)$ contains only the zero element. If z is on the boundary, then $C(z)$ is the infinite convex cone generated by the outer normal at z of the faces on which z lies; see [17, section 4.3] for more discussions. The proof of the proposition is based on a combined use of a probabilistic approach and analytic results on differential equations. For explanations on the conditions needed together with a proof, we refer the reader to [23]. In addition to the convergence, we may also study the rates of convergence and obtain large deviation-type bounds as was done in [25]. However, these are not the main concerns of the current paper. We are more interested in the numerical performance of the algorithm, which is discussed next.

4.2. Numerical results.

In this section we provide two numerical examples and compare the results. We study a two-dimensional problem with variable parameters (i.e., regime-dependent parameters) in the first example and constant parameters in

Table 1
Optimal thresholds using the stochastic approximation algorithm.

Initial z	$-0.05, 0.05$	$-0.10, 0.20$	$-0.20, 0.40$	$-0.10, 0.60$	$-0.05, 0.85$
$z^*(n_0 = 100)$	$-0.36, 0.421$	$-0.36, 0.422$	$-0.36, 0.422$	$-0.36, 0.422$	$-0.36, 0.423$
$z^*(n_0 = 10)$	$-0.36, 0.409$	$-0.36, 0.418$	$-0.36, 0.429$	$-0.36, 0.418$	$-0.36, 0.419$
$z^*(n_0 = 1)$	$-0.36, 0.409$	$-0.36, 0.416$	$-0.36, 0.425$	$-0.359, 0.417$	$-0.359, 0.423$

the second. In both cases, the Markov chain $\alpha(t)$ takes two states, whose generator is given by

$$Q = \begin{pmatrix} -6.04 & 6.04 \\ 8.90 & -8.90 \end{pmatrix}.$$

The probability distribution of the initial Markov chain $\alpha(0)$ is given by $p_1 = p_2 = \frac{1}{2}$. We use the utility function $\Phi(x) = e^x - 1$ (see Remark 2.4).

Example 4.3. We choose the following parameter values for the regime-switching model: $\mu(1) = 0.01$, $\mu(2) = 0.02$, $b(1) = b(2) = 0$, $\sigma^2(1) = 0.25$, $\sigma^2(2) = 0.5$, and $\rho = 1$. We first implement the stochastic recursive algorithm developed in section 4.1. For the search region for $z = (z_1, z_2)$, we choose $(z_1, z_2) \in [a_1, b_1] \times [a_2, b_2] = [-0.36, -0.01] \times [0.01, 1.0]$. The sequence $\{\varepsilon_n\}$ for step sizes in (4.4) and the sequence $\{\delta_n\}$ used in the gradient estimation (4.3) are chosen to be $\varepsilon_n = 1/(n + k_0)$ and $\delta_n = 1/(n^{1/6} + k_1)$, respectively, where k_0 and k_1 are some positive integers, e.g., $k_0 = k_1 = 1$. The search stops whenever $\varepsilon_n < 0.001$. In what follows, we use n_0 replications, as presented in Remark 4.1. Table 1 reports the search results by using the stochastic recursive algorithm for five different initial values of z and for three different n_0 for gradient estimation. Note that the last row in the table ($n_0 = 1$) gives the results obtained by using a single path gradient estimate in the recursion.

Next we numerically solve the differential equations (2.11) with boundary conditions (2.12). For this example, they become

$$\text{(4.10)} \qquad \begin{cases} \dfrac{\sigma^2(1)}{2}\dfrac{d^2v(z,1)}{dz^2} + \mu(1)z\dfrac{dv(z,1)}{dz} + (q_{11} - \rho)v(z,1) + q_{12}v(z,2) = 0, \\ \dfrac{\sigma^2(2)}{2}\dfrac{d^2v(z,2)}{dz^2} + \mu(2)z\dfrac{dv(z,2)}{dz} + (q_{22} - \rho)v(z,2) + q_{21}v(z,1) = 0, \\ v(z_1, i) = e^{z_1} - 1, \quad v(z_2, i) = e^{z_2} - 1, \quad i = 1, 2. \end{cases}$$

We use a grid size 0.01 to divide the region $[-0.36, -0.01] \times [0.01, 1.0]$ for (z_1, z_2). This results in 36 points along z_1, 100 points along z_2, and totally 3600 different pairs for (z_1, z_2). For each pair, which specifies the boundary values, a finite difference scheme is used to solve the system (4.10). The objective function $V(z_1, z_2)$ is then calculated by $V(z_1, z_2) = [v(0,1) + v(0,2)]/2$. Figure 2 plots the surface $V(z_1, z_2)$ using the 3600 values. The numerical results show that the maximum value for $V(z_1, z_2)$ is achieved at $(-0.36, 0.41)$ and $(-0.36, 0.42)$. This suggests that the optimal threshold (z_1^*, z_2^*) is very close to these two points. It is consistent with the estimates obtained in Table 1 by using the stochastic optimization algorithms. Note that numerically solving the differential equations is time consuming, while the stochastic recursive algorithms produce the optimal estimates in much less computation. This efficiency becomes more eminent when a small number of sample paths is used in gradient estimation. From Table 1 we notice that even a single sample path yields pretty good approximations to the optimal thresholds.

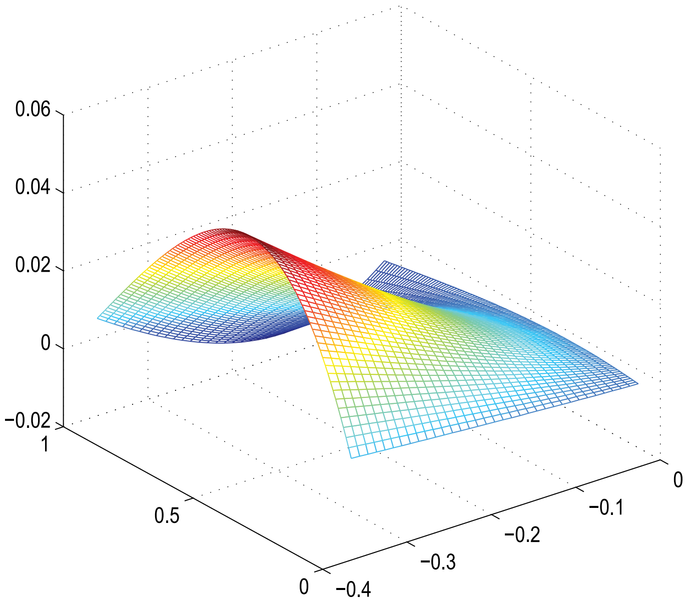


FIG. 2. *Surface of the value function* $V(z_1, z_2)$ *over the region* $(z_1, z_2) \in [-0.36, -0.01] \times [0.01, 1.0]$. *Grid size* 0.01 *is used.*

TABLE 2
Comparison of optimal selling rules in different markets.

	Optimal threshold (z_1^*, z_2^*)	Percentage increase in asset price	Percentage decrease in asset price
Bear market (Case I, Ex. 3)	$(-0.36, 0.33)$	39%	30%
Bull market (Case II, Ex. 3)	$(-0.36, 0.55)$	73%	30%
Mixed market (Ex. 2)	$(-0.36, 0.42)$	52%	30%

Based on the results, we may conclude that the optimal threshold for this specific example is given by $(z_1^*, z_2^*) = (-0.36, 0.42)$ with double-digit precision. This pair of values corresponds to a 52% increase and a 30% decrease in asset price, respectively. Following the selling rule, an investor would sell the asset he or she has bought whenever the price goes up by 52% or down by 30%.

Example 4.4. In this example we assume that the model parameters do not change across regimes, i.e., $\mu(1) = \mu(2) = \mu$, $\sigma(1) = \sigma(2) = \sigma$, while keeping other values the same, as in the last example. We report two cases: one uses regime 1 parameters and another uses regime 2 parameters from Example 4.3.

Case 1. $\mu = 0.01$, $\sigma^2 = 0.25$. The optimal thresholds are $(z_1^*, z_2^*) = (-0.36, 0.33)$, which correspond to a 39% increase and a 30% decrease in asset price.

Case 2. $\mu = 0.02$, $\sigma^2 = 0.50$. The optimal thresholds are $(z_1^*, z_2^*) = (-0.36, 0.55)$, which correspond to a 73% increase and a 30% decrease in asset price.

For comparison, in Table 2, we summarize the results from Examples 4.4 and 4.3. We may call Case 2 in Example 4.4 a bull market since a bigger μ value is used and Case 1 a bear market since a smaller μ value is used. Then we call Example 4.3 a mixed market because of the switching between the two μ numbers. Note that the optimal selling rules change in a manner that agrees with common investment practice. If a 30% drop in asset price is used by investors for the stop-loss limit, then the upper threshold for achieving maximum profit is higher (73%) in the bull market than that in the (more realistic) mixed market (52%), which in turn is higher than that in the bear market (39%).

5. Concluding remarks. In this paper we developed an optimal selling rule using a regime-switching exponential Gaussian diffusion model. The optimal selling can be characterized by two threshold levels. We designed a numerical algorithm for searching these threshold levels.

Note that our results in this paper rely crucially on Assumption 3.4. It is interesting and practically useful to relax these conditions. In addition, we assumed the market mode to be completely observable. In order to apply our results in practice, one needs to estimate the system mode using nonlinear filtering techniques. The Wonham filter, in which the hidden Markov chain $\alpha(t)$ is observed in noise, is a good candidate; it provides sound conditional probability estimates given the stock price up to time t.

Acknowledgment. We thank the reviewers whose comments have improved the exposition of this paper.

REFERENCES

[1] R. Bansal and H. Zhou, *Term structure of interest rates with regime shifts*, J. Finance, 57 (2002), pp. 1997–2043.

[2] J. Buffington and R. J. Elliott, *American options with regime switching*, Int. J. Theor. Appl. Finance, 5 (2002), pp. 497–514.

[3] L. Clewlow and C. Strickland, *Energy Derivatives: Pricing and Risk Management*, Lacima Publications, London, 2000.

[4] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[5] W. A. Coppel, *Disconjugacy*, Lecture Notes in Math. 220, Springer-Verlag, New York, 1971.

[6] Q. Dai, K. J. Singleton, and W. Yang, *Regime Shifts in a Dynamic Term Structure Model of U.S. Treasury Bond Yields*, working paper, 2005.

[7] D. A. Darling and A. J. F. Siegert, *The first passage problem for a continuous Markov process*, Ann. Math. Statistics, 24 (1953), pp. 624–639.

[8] D. Duffie, D. Filipović, and W. Schachermayer, *Affine processes and applications in finance*, Ann. Appl. Probab., 13 (2003), pp. 984–1053.

[9] C. Erlwein, F. Benth, and R. Mamon, *HMM Filtering and Parameter Estimation of an Electricity Spot Price Model*, E-print 2, Dept. of Mathematics, University of Oslo, Norway, 2007.

[10] S. Finch, *Ornstein-Uhlenbeck Process*, working notes, 2004.

[11] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York, 2004.

[12] X. Guo, *Information and option pricings*, Quant. Finance, 1 (2001), pp. 38–44.

[13] X. Guo and Q. Zhang, *Closed-form solutions for perpetual American put options with regime switching*, SIAM J. Appl. Math., 64 (2004), pp. 2034–2049.

[14] L. K. Jackson, *Boundary value problems for ordinary differential equations*, in Studies in Ordinary Differential Equations, MAA Studies in Mathematics 14, J. K. Hale, ed., Mathematical Association of America, Washington, DC, 1977.

[15] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.

[16] T. Kluge, *Pricing Swing Options and Other Electricity Derivatives*, Ph.D./D.Phil. thesis, University of Oxford, Oxford, UK, 2006.

[17] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.

[18] J. Lucia and E. Schwartz, *Electricity prices and power derivatives: Evidence from the Nordic power exchange*, Review of Derivatives Research, 5 (2002), pp. 5–50.

[19] B. K. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, New York, 2003.

[20] J. M. Steele, *Stochastic Calculus and Financial Applications*, Springer-Verlag, New York, 2003.

[21] O. Vasicek, *An equilibrium characterization of the term structure*, J. Financial Economics, 5 (1977), pp. 177–188.

[22] D. D. Yao, Q. Zhang, and X. Y. Zhou, *A regime-switching model for European options*, in Stochastic Processes, Optimization, and Control Theory: Applications in Financial Engineering, Queueing Networks, and Manufacturing Systems, H. M. Yan, G. Yin, and Q. Zhang, eds., Springer-Verlag, New York, 2006, pp. 281–300.

[23] G. Yin, R. H. Liu, and Q. Zhang, *Recursive algorithms for stock liquidation: A stochastic optimization approach*, SIAM J. Optim., 13 (2002), pp. 240–263.

[24] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.

[25] G. Yin, Q. Zhang, F. Liu, R. H. Liu, and Y. Cheng, *Stock liquidation via stochastic approximation using NASDAQ daily and intra-day data*, Math. Finance, 16 (2006), pp. 217–236.

[26] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.

[27] D. Zwillinger, *Handbook of Differential Equations*, Academic Press, New York, 1989.

SIAM J. APPL. MATH.
Vol. 69, No. 3, pp. 830–844

FAR FIELD MODELING OF ELECTROMAGNETIC TIME REVERSAL AND APPLICATION TO SELECTIVE FOCUSING ON SMALL SCATTERERS*

X. ANTOINE†, B. PINÇON†, K. RAMDANI†, AND B. THIERRY†

Abstract. A time harmonic far field model for closed electromagnetic time reversal mirrors is proposed. Then, a limit model corresponding to small perfectly conducting scatterers is derived. This asymptotic model is used to prove the selective focusing properties of the time reversal operator. In particular, a mathematical justification of the decomposition of the time reversal operator (DORT) method is given for axially symmetric scatterers.

Key words. electromagnetic scattering, time reversal, far field, small obstacles

AMS subject classifications. 35B40, 35P25, 45A05, 74J20, 78M35

DOI. 10.1137/080715779

1. Introduction.

In the last decade, acoustic time reversal has definitely demonstrated its efficiency in target characterization by wave focusing in complex media (see the review papers [13, 15]). In particular, it has been shown that selective focusing can be achieved using the eigenvectors (resp., eigenfunctions) of the so-called time reversal matrix (resp., operator). Known as the DORT method (French acronym for *diagonalization of the time reversal operator*; see [14, 32, 26, 31, 16, 25, 18]), this technique involves three steps. First, an incident wave is emitted in the medium containing the scatterers by the time reversal mirror (TRM). The scattered field is then measured by the mirror and time-reversed (or phase-conjugated in the time harmonic case). Finally, the obtained signal is then re-emitted in the medium. By definition, the time reversal operator $\mathbf{T}$ is the operator describing two successive cycles of emission/reception/time reversal. If the propagation medium is nondissipative, then the operator $\mathbf{T}$ is hermitian, since $\mathbf{T} = \mathbf{F}^*\mathbf{F}$, where $\mathbf{F}$ denotes the far field operator. The DORT method can thus be seen as a singular value decomposition of $\mathbf{F}$. Moreover, in a particular range of frequencies (for which the scatterers can be considered as point-like scatterers), $\mathbf{T}$ has as many significant eigenvalues as there are scatterers in the medium, and the corresponding eigenfunctions generate incident waves that selectively focus on the scatterers. From the mathematical point of view, a detailed analysis of this problem has been proposed for the acoustic scattering problem by small scatterers in the free space in [19] and in a two-dimensional straight waveguide in [29]. Let us emphasize that time reversal has also been intensively studied in the context of random media (see [17] and the references therein).

Recently, electromagnetic focusing using time reversal has been demonstrated experimentally [23] and used for imaging applications [24]. One of the first works dealing with mathematical and numerical aspects of electromagnetic time reversal is the paper [34]. The authors analyze therein the DORT method in the case of a homogeneous

*Received by the editors February 15, 2008; accepted for publication (in revised form) September 10, 2008; published electronically December 19, 2008.
http://www.siam.org/journals/siap/69-3/71577.html

†Institut Elie Cartan Nancy (Nancy-Université, CNRS, INRIA), Université Henri Poincaré, BP 239, 54506, Vandœuvre-lès-Nancy, France, and INRIA (Corida Team), 615 rue du Jardin Botanique, 54600 Villers-lès-Nancy, France (Xavier.Antoine@iecn.u-nancy.fr, Bruno.Pincon@iecn.u-nancy.fr, Karim.ramdani@loria.fr, Bertrand.Thierry@iecn.u-nancy.fr).

medium containing perfectly conducting or dielectric objects of particular shapes (circular rods and spheres). Their method is based on a low frequency approximation of a multipole expansion of the scattered field (i.e., a Fourier–Bessel series involving Hankel functions for circular rods and vector spherical functions for spheres). In [8], the authors proposed an iterative process based on time reversal to determine optimal electromagnetic measurements (i.e., to determine the incident waves maximizing the scattered field). More recently, the DORT method has been used for target localization, especially in the context of imaging [6, 7, 1]. The analysis followed in these works is based on the singular value decomposition of the multistatic response matrix, which corresponds to the case where the mirror is described by a discrete array of transducers (emitters and receivers). In this paper, we propose a time harmonic far field model of electromagnetic time reversal in the case of a continuous distribution of transducers. Only *closed mirrors* (i.e., completely surrounding the scatterers) are considered in this work, and the limited aperture case is not studied. Except for this difference, the present work can be seen as an extension of the results obtained for acoustic time reversal in free space [19] and in straight waveguides [29]. We pay very careful attention to the derivation of the limit scattering model for small perfectly conducting scatterers. The functional framework used hereafter for the far field and the time reversal operators is the one commonly used in inverse electromagnetic scattering theory [11, 5, 20].

We start the paper with a short description in section 2 of the mathematical model of time reversal. In particular, we define the incident field emitted by the TRM (electromagnetic Herglotz waves), the measured fields (the far field pattern), and the time reversal operator. In section 3, we restrict our analysis to the case of small scatterers (of typical size δ). We show that the small scatterers' asymptotics can be deduced from the classical low frequency scattering asymptotics (the Rayleigh approximation) involving the polarization tensors of the scatterers. More precisely, our analysis corresponds to the case where $k\delta$ and δ/d tend simultaneously to 0, where k denotes the wavenumber and d the minimum separation distance between the scatterers. Finally, we study in section 4 the spectral focusing properties of the eigenfunctions of the limit far field operator obtained in section 3. We show that each small scatterer gives rise to at most six distinct eigenvalues (recovering the results obtained in [7, 1] for the case of a discrete TRM). Furthermore, if the polarizability tensors of the scatterers are diagonal (e.g., for axially symmetric scatterers) and under the additional assumption that $kd \to \infty$, we prove that each associated eigenfunction generates an incident wave that selectively focuses on the corresponding scatterer.

2. A far field model for electromagnetic time reversal. In order to obtain an expression of the time reversal operator, we begin this paper by recalling the far field model of electromagnetic scattering. Consider the scattering problem of an incident electromagnetic plane wave by a perfectly conducting bounded obstacle contained in a homogeneous medium. Without loss of generality, we assume that the electric permittivity ε and the magnetic permeability μ are both equal to 1. Let $\mathcal{O}$ be a bounded open subset of $\mathbb{R}^3$ with smooth boundary Γ and outward unit normal $\boldsymbol{\nu}$, and let $\Omega = \mathbb{R}^3 \setminus \overline{\mathcal{O}}$ be the propagation domain. Let $L^2_t(S^2)$ be the space of tangential vector fields of the unit sphere S^2,

$$L^2_t(S^2) = \left\{ \boldsymbol{f} \in \left(L^2(S^2)\right)^3 \mid \forall \boldsymbol{\alpha} \in S^2,\, \boldsymbol{f}(\boldsymbol{\alpha}) \cdot \boldsymbol{\alpha} = 0 \right\},$$

and consider the incident plane wave $(\boldsymbol{E}_I^{\boldsymbol{\alpha},\boldsymbol{f}}, \boldsymbol{H}_I^{\boldsymbol{\alpha},\boldsymbol{f}})$ of direction $\boldsymbol{\alpha} \in S^2$ and electric polarization $\boldsymbol{f} \in L_t^2(S^2)$:

$$\left\{ \begin{array}{l} \boldsymbol{E}_I^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{\alpha})\, \mathrm{e}^{ik\boldsymbol{\alpha}\cdot\boldsymbol{x}}, \\ \boldsymbol{H}_I^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) = (\boldsymbol{\alpha} \times \boldsymbol{f}(\boldsymbol{\alpha}))\, \mathrm{e}^{ik\boldsymbol{\alpha}\cdot\boldsymbol{x}}. \end{array} \right. \tag{2.1}$$

Throughout the paper, the time dependence is assumed to be of the form $e^{-i\omega t}$ and will always be implicit. Introducing the wavenumber $k = \omega\sqrt{\varepsilon\mu} = \omega$, the scattered field $(\boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}})$ solves the following exterior boundary value problem:

$$\begin{cases} \operatorname{curl} \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}} = ik\, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}} & (\Omega), \\ \operatorname{curl} \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}} = -ik\, \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}} & (\Omega), \\ \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}} \times \boldsymbol{\nu} = -\boldsymbol{E}_I^{\boldsymbol{\alpha},\boldsymbol{f}} \times \boldsymbol{\nu} & (\Gamma), \\ \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}} \cdot \boldsymbol{\nu} = -\boldsymbol{H}_I^{\boldsymbol{\alpha},\boldsymbol{f}} \cdot \boldsymbol{\nu} & (\Gamma), \\ \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}} \text{ are outgoing.} \end{cases} \tag{2.2}$$

Classically, the outgoing behavior of the scattered field is imposed by one of the two Silver–Müller radiation conditions,

$$\begin{cases} \lim\limits_{|\boldsymbol{x}|\to\infty} \Big(\boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) \times \boldsymbol{x} + |\boldsymbol{x}|\, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) \Big) = 0, \\ \lim\limits_{|\boldsymbol{x}|\to\infty} \Big(\boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) \times \boldsymbol{x} - |\boldsymbol{x}|\, \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x}) \Big) = 0, \end{cases}$$

uniformly in every direction $\boldsymbol{x}/|\boldsymbol{x}| \in S^2$, where $|\,.\,|$ is the Euclidean norm in $\mathbb{R}^3$.

We are now in a position to introduce the far field pattern of the electromagnetic field $(\boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}})$. Its main properties are collected in the next proposition (see [11] for the proofs).

Proposition 2.1. *The scattered field* $(\boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}})$ *has the following asymptotic behavior in the direction* $\boldsymbol{\beta} \in S^2$ *as* $|\boldsymbol{x}| \to \infty$*:*

$$\left\{ \begin{array}{l} \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{\beta}|\boldsymbol{x}|) = \dfrac{e^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|} \boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) + O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), \\ \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{\beta}|\boldsymbol{x}|) = \dfrac{e^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|} (\boldsymbol{\beta} \times \boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))) + O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right). \end{array} \right.$$

The scattering amplitude $\boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))$ *is given for all* $\boldsymbol{\alpha},\boldsymbol{\beta} \in S^2$ *and all* $\boldsymbol{f} \in L_t^2(S^2)$ *by the formula*

$$\boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) = -\frac{k^2}{4\pi}\boldsymbol{\beta} \times \int_\Gamma \Big[\boldsymbol{\nu}(\boldsymbol{y}) \times \boldsymbol{H}_T^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{y})\Big] \times \boldsymbol{\beta}\, e^{-ik\boldsymbol{\beta}\cdot\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y}, \tag{2.3}$$

where $\boldsymbol{H}_T^{\boldsymbol{\alpha},\boldsymbol{f}} = \boldsymbol{H}_I^{\boldsymbol{\alpha},\boldsymbol{f}} + \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}}$ *is the total magnetic field. Moreover,* $\boldsymbol{A}(\cdot,\cdot;\cdot)$ *satisfies the reciprocity relation*

$$\boldsymbol{g}(\boldsymbol{\beta}) \cdot \boldsymbol{A}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) = \boldsymbol{f}(\boldsymbol{\alpha}) \cdot \boldsymbol{A}(-\boldsymbol{\beta},-\boldsymbol{\alpha};\boldsymbol{g}(\boldsymbol{\beta})) \tag{2.4}$$

for all $\boldsymbol{\alpha},\boldsymbol{\beta} \in S^2$ *and all* $\boldsymbol{f},\boldsymbol{g} \in L_t^2(S^2)$.

Assume now that the TRM emits a Herglotz wave, i.e., a superposition of plane waves of the form (2.1). More precisely, denote by $(\boldsymbol{E}_I^{\boldsymbol{f}}, \boldsymbol{H}_I^{\boldsymbol{f}})$ the incident Herglotz wave of polarization $\boldsymbol{f} \in L^2_t(S^2)$, defined by

$$\begin{cases} \boldsymbol{E}^{\boldsymbol{f}}_I(\boldsymbol{x}) = \displaystyle\int_{S^2} \boldsymbol{E}^{\boldsymbol{\alpha},\boldsymbol{f}}_I(\boldsymbol{x})\,\mathrm{d}\boldsymbol{\alpha} = \int_{S^2} \boldsymbol{f}(\boldsymbol{\alpha}) e^{ik\boldsymbol{\alpha}\cdot\boldsymbol{x}}\,\mathrm{d}\boldsymbol{\alpha}, \\ \boldsymbol{H}^{\boldsymbol{f}}_I(\boldsymbol{x}) = \displaystyle\int_{S^2} \boldsymbol{H}^{\boldsymbol{\alpha},\boldsymbol{f}}_I(\boldsymbol{x})\,\mathrm{d}\boldsymbol{\alpha} = \int_{S^2} (\boldsymbol{\alpha} \times \boldsymbol{f})(\boldsymbol{\alpha}) e^{ik\boldsymbol{\alpha}\cdot\boldsymbol{x}}\,\mathrm{d}\boldsymbol{\alpha}. \end{cases} \tag{2.5}$$

By linearity, Proposition 2.1 yields the following result.

COROLLARY 2.2. *When illuminated by the Herglotz wave* $(\boldsymbol{E}^{\boldsymbol{f}}_I, \boldsymbol{H}^{\boldsymbol{f}}_I)$, *the scattering obstacle generates the diffracted field* $(\boldsymbol{E}^{\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{f}})$, *which admits in the direction* $\boldsymbol{\beta} \in S^2$ *the far field asymptotics*

$$\begin{cases} \boldsymbol{E}^{\boldsymbol{f}}(\boldsymbol{\beta}|\boldsymbol{x}|) = \dfrac{e^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|}\mathbf{F}\boldsymbol{f}(\boldsymbol{\beta}) + O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), \\ \boldsymbol{H}^{\boldsymbol{f}}(\boldsymbol{\beta}|\boldsymbol{x}|) = \dfrac{e^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|}\boldsymbol{\beta} \times \mathbf{F}\boldsymbol{f}(\boldsymbol{\beta}) + O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), \end{cases}$$

where $\mathbf{F}\boldsymbol{f}(\boldsymbol{\beta})$ *is given by*

$$\mathbf{F}\boldsymbol{f}(\boldsymbol{\beta}) = \int_{S^2} \boldsymbol{A}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{f}(\boldsymbol{\alpha}))\,\mathrm{d}\boldsymbol{\alpha}. \tag{2.6}$$

Using the expression (2.3) of the scattering amplitude, one can show that the far field operator $\mathbf{F} : \boldsymbol{f} \longmapsto \mathbf{F}\boldsymbol{f}$ defined by (2.6) is continuous from $L^2_t(S^2)$ onto itself. Moreover, using the reciprocity relation (2.4), one can show the following result (see [9] for the proof).

PROPOSITION 2.3. *The far field operator* $\mathbf{F} : L^2_t(S^2) \longrightarrow L^2_t(S^2)$ *defined by* (2.6) *is a compact and normal operator. As in the acoustic case, its adjoint is the operator* $\mathbf{F}^* : L^2_t(S^2) \to L^2_t(S^2)$ *defined by*

$$\forall \boldsymbol{f} \in L^2_t(S^2), \qquad \mathbf{F}^*\boldsymbol{f} = \overline{\mathbf{R}\mathbf{F}\overline{\mathbf{R}\boldsymbol{f}}}, \tag{2.7}$$

where $\mathbf{R}$ *is the symmetry operator defined by* $\mathbf{R}\boldsymbol{f}(\boldsymbol{\alpha}) = \boldsymbol{f}(-\boldsymbol{\alpha})$ *for all* $\boldsymbol{\alpha} \in S^2$ *and* $\boldsymbol{f} \in L^2_t(S^2)$.

We are now able to define the time reversal operator $\mathbf{T}$. During the time reversal process, the TRM first emits an incident electromagnetic Herglotz wave $(\boldsymbol{E}^{\boldsymbol{f}}_I, \boldsymbol{H}^{\boldsymbol{f}}_I)$ of polarization $\boldsymbol{f}$. Then the scattering obstacle generates a scattered field $(\boldsymbol{E}^{\boldsymbol{f}}, \boldsymbol{H}^{\boldsymbol{f}})$. The TRM measures and conjugates the corresponding electric far field $\mathbf{F}\boldsymbol{f}$. The resulting field is then used as a polarization $\boldsymbol{g}$ of a new incident Herglotz wave. Therefore, we have

$$\boldsymbol{g} = \overline{\mathbf{R}\mathbf{F}\boldsymbol{f}}, \tag{2.8}$$

where the presence of the symmetry operator is due to the fact that the far field measured in a direction $\boldsymbol{\beta}$ is re-emitted in the opposite direction $-\boldsymbol{\beta}$. The time reversal operator $\mathbf{T}$ is then obtained by iterating this cycle twice:

$$\mathbf{T}\boldsymbol{f} = \overline{\mathbf{R}\mathbf{F}\boldsymbol{g}} = \overline{\mathbf{R}\mathbf{F}\overline{\mathbf{R}\mathbf{F}\boldsymbol{f}}}. \tag{2.9}$$

Thanks to Proposition 2.3, we have shown the following result.

PROPOSITION 2.4. *The time reversal operator* $\mathbf{T}$ *is the compact, self-adjoint, and positive operator given by*

$$\begin{array}{rlcl} \mathbf{T}: & L^2_t(S^2) & \longrightarrow & L^2_t(S^2), \\ & \boldsymbol{f} & \longmapsto & \mathbf{T}\boldsymbol{f} = \mathbf{F}\mathbf{F}^*\boldsymbol{f} = \mathbf{F}^*\mathbf{F}\boldsymbol{f}. \end{array}$$

The nonzero eigenvalues of $\mathbf{T}$ *are exactly the positive numbers*

$$|\lambda_1|^2 \geq |\lambda_2|^2 \geq \cdots > 0,$$

where the sequence $(\lambda_p)_{p\geq 1}$ *denotes the nonzero complex eigenvalues of the normal compact far field operator* $\mathbf{F}$*. Moreover, the corresponding eigenfunctions* $(\boldsymbol{f}_p)_{p\geq 1}$ *of* $\mathbf{F}$ *are exactly the eigenfunctions of* $\mathbf{T}$*.*

3. Scattering by perfectly conducting small scatterers.

In this section, we show that the asymptotics of the electromagnetic scattering problem by small scatterers is closely connected to the classical low frequency scattering (the Rayleigh approximation [21, 12]). In particular, this asymptotics involves the electromagnetic polarizability tensors of the scatterers [30, 2, 3]. The fact that the two limit models are similar is straightforward when the scattering obstacle has only one connected component. As shown in subsection 3.1, this follows from a scaling argument. The proof is less obvious when the obstacle is multiply connected (one can no longer use a unique change of variables to work in a reference domain of fixed size). We study this question using an integral equation approach in subsection 3.2.

3.1. The case of one scatterer.

Let us assume that the perfectly conducting scatterer is of small size δ and that it is obtained from a reference obstacle after a dilation. More precisely, let us set

$$\mathcal{O}^\delta = \{\boldsymbol{x} = \boldsymbol{s} + \delta\boldsymbol{\xi}\ ;\ \boldsymbol{\xi} \in \mathcal{O}\}.$$

Its boundary is denoted by Γ^δ and its exterior by $\Omega^\delta := \mathbb{R}^3 \setminus \overline{\mathcal{O}^\delta}$. Given an incident plane wave $(\boldsymbol{E}_I^{\alpha,\boldsymbol{f}}, \boldsymbol{H}_I^{\alpha,\boldsymbol{f}})$, let $(\boldsymbol{E}^\delta, \boldsymbol{H}^\delta)$ be the solution of the scattering problem by the perfectly conducting obstacle (for the sake of clarity, we drop here the reference to the angle of incidence and to the polarization in the scattered field):

$$\left\{\begin{array}{ll}\operatorname{curl}\boldsymbol{E}^\delta = ik\boldsymbol{H}^\delta & (\Omega^\delta),\\ \operatorname{curl}\boldsymbol{H}^\delta = -ik\boldsymbol{E}^\delta & (\Omega^\delta),\\ \operatorname{div}\boldsymbol{E}^\delta = 0 & (\Omega^\delta),\\ \operatorname{div}\boldsymbol{H}^\delta = 0 & (\Omega^\delta),\\ \boldsymbol{E}^\delta \times \boldsymbol{\nu} = -\boldsymbol{E}_I^{\alpha,\boldsymbol{f}} \times \boldsymbol{\nu} & (\Gamma^\delta),\\ \boldsymbol{H}^\delta \cdot \boldsymbol{\nu} = -\boldsymbol{H}_I^{\alpha,\boldsymbol{f}} \cdot \boldsymbol{\nu} & (\Gamma^\delta),\\ \boldsymbol{E}^\delta,\ \boldsymbol{H}^\delta \text{ outgoing.} & \end{array}\right. \tag{3.1}$$

Introducing the scaled fields

$$\left\{\begin{array}{l}\boldsymbol{e}^\delta(\boldsymbol{\xi}) = \boldsymbol{E}^\delta(\boldsymbol{s} + \delta\boldsymbol{\xi}),\\ \boldsymbol{h}^\delta(\boldsymbol{\xi}) = \boldsymbol{H}^\delta(\boldsymbol{s} + \delta\boldsymbol{\xi}),\end{array}\right. \qquad \boldsymbol{\xi} \in \Omega := \mathbb{R}^3 \setminus \overline{\mathcal{O}},$$

we obtain that

$$\left\{\begin{array}{ll}\operatorname{curl}\boldsymbol{e}^\delta = i\,(k\delta)\,\boldsymbol{h}^\delta & (\Omega),\\ \operatorname{curl}\boldsymbol{h}^\delta = -i\,(k\delta)\,\boldsymbol{e}^\delta & (\Omega),\\ \operatorname{div}\boldsymbol{e}^\delta = 0 & (\Omega),\\ \operatorname{div}\boldsymbol{h}^\delta = 0 & (\Omega),\\ \boldsymbol{e}^\delta \times \boldsymbol{\nu} = -\boldsymbol{e}_I^{\alpha,\boldsymbol{f}} \times \boldsymbol{\nu} & (\Gamma),\\ \boldsymbol{h}^\delta \cdot \boldsymbol{\nu} = -\boldsymbol{h}_I^{\alpha,\boldsymbol{f}} \cdot \boldsymbol{\nu} & (\Gamma),\\ \boldsymbol{e}^\delta,\ \boldsymbol{h}^\delta \text{ outgoing,} & \end{array}\right. \tag{3.2}$$

where $\Gamma = \partial\Omega$ and

$$\begin{cases} \boldsymbol{e}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{\xi}) = \boldsymbol{E}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}+\delta\boldsymbol{\xi}) = \boldsymbol{E}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}) + O(k\delta), \\ \boldsymbol{h}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{\xi}) = \boldsymbol{H}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}+\delta\boldsymbol{\xi}) = \boldsymbol{H}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}) + O(k\delta). \end{cases}$$

When $\delta \to 0$, problem (3.2) appears as a low frequency electromagnetic scattering problem ($k\delta \to 0$) associated with an incident wave that behaves like the constant field $(\boldsymbol{E}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}), \boldsymbol{H}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}))$ asymptotically. The electromagnetic scattering problem for small frequencies has been studied for a long time (see [35, 36, 22, 28]), and the asymptotic behavior of its solution is by now well known (see the reference book [12] for a detailed presentation and [4] for convergence results of higher order terms). In particular, the first order approximation $(\boldsymbol{e}^0, \boldsymbol{h}^0)$ of $(\boldsymbol{e}^\delta, \boldsymbol{h}^\delta)$ (the so-called Rayleigh approximation) is given by the next result, which follows from [12, Chap. 5]).

THEOREM 3.1. *Let $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \Phi_3)$ and $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, \Psi_3)$ be the vector potentials defined by*

$$\text{(3.3)} \qquad \begin{cases} \Delta\boldsymbol{\Phi} = 0 & (\Omega), \\ \boldsymbol{\Phi} = \boldsymbol{x} + \boldsymbol{c} & (\Gamma), \\ \boldsymbol{\Phi} = O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), & |\boldsymbol{x}| \to \infty, \end{cases}$$

and

$$\text{(3.4)} \qquad \begin{cases} \Delta\boldsymbol{\Psi} = 0 & (\Omega), \\ \dfrac{\partial\boldsymbol{\Psi}}{\partial\boldsymbol{\nu}} = \boldsymbol{\nu} & (\Gamma), \\ \boldsymbol{\Psi} = O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), & |\boldsymbol{x}| \to \infty, \end{cases}$$

where the constant vector $\boldsymbol{c} \in \mathbb{R}^3$ is chosen such that $\int_\Gamma \frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\nu}} = 0$.

Then, as $\delta \longrightarrow 0$, we have

$$\begin{cases} \boldsymbol{e}^\delta \longrightarrow \boldsymbol{e}^0 := -\nabla\boldsymbol{\Phi}\, \boldsymbol{f}(\boldsymbol{\alpha}), \\ \boldsymbol{h}^\delta \longrightarrow \boldsymbol{h}^0 := -\, \nabla\boldsymbol{\Psi}\, (\boldsymbol{\alpha} \times \boldsymbol{f}(\boldsymbol{\alpha})) \end{cases}$$

locally in $H_{\mathrm{curl}}(\Omega)$.

Using the above result, one can easily obtain the asymptotics of the far field associated to $\boldsymbol{E}^\delta$.

COROLLARY 3.2. *Let $(\boldsymbol{E}^\delta, \boldsymbol{H}^\delta)$ be the solution of the scattering problem (3.1). Let $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\mathcal{M}}$ be, respectively, the electric polarizability and magnetic polarizability tensors defined by (I denotes the identity)*

$$\boldsymbol{\mathcal{P}} = |\mathcal{O}|\, I - \int_\Gamma \boldsymbol{x} \left(\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\nu}}\right)^T \mathrm{d}\gamma_{\boldsymbol{x}}, \qquad \boldsymbol{\mathcal{M}} = |\mathcal{O}|\, I - \int_\Gamma \boldsymbol{\nu}\boldsymbol{\Psi}^T \,\mathrm{d}\gamma_{\boldsymbol{x}},$$

where the vector potentials $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are respectively defined by (3.3) *and* (3.4). *Then, the far field $\boldsymbol{A}^\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{f}(\boldsymbol{\alpha}))$ of $\boldsymbol{E}^\delta$, defined by*

$$\boldsymbol{E}^\delta(\boldsymbol{\beta}|\boldsymbol{x}|) = \boldsymbol{A}^\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{f}(\boldsymbol{\alpha})) \frac{\mathrm{e}^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|} + O\left(\frac{1}{|\boldsymbol{x}|^2}\right),$$

admits as $\delta \to 0$ the following asymptotics:

$$\text{(3.5)} \quad \boldsymbol{A}^\delta(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{f}(\boldsymbol{\alpha})) = \frac{(ik\delta)^3}{4\pi}\, \boldsymbol{\beta} \times \Big[\boldsymbol{\beta} \times (\boldsymbol{\mathcal{P}}\boldsymbol{f}(\boldsymbol{\alpha})) - \boldsymbol{\mathcal{M}}(\boldsymbol{\alpha} \times \boldsymbol{f}(\boldsymbol{\alpha}))\Big]\, \mathrm{e}^{ik\,(\boldsymbol{\alpha}-\boldsymbol{\beta})\cdot\boldsymbol{s}} + O(\delta^4).$$

Proof. Following [12], we have

$$\boldsymbol{A}^{\delta}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) = \frac{k^2}{4\pi}\,\boldsymbol{\beta}\times\left\{\boldsymbol{\beta}\times\int_{\Gamma^{\delta}}\left[\boldsymbol{\nu}_{\boldsymbol{x}}\times\left(\boldsymbol{H}^{\delta}(\boldsymbol{x})+\boldsymbol{H}_I^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{x})\right)\right]\,\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{x}}\,\mathrm{d}\gamma_{\boldsymbol{x}}\right\}. \tag{3.6}$$

The change of variables $\boldsymbol{\xi} = (\boldsymbol{x}-\boldsymbol{s})/\delta$ in the above integral shows that

$$\begin{aligned}\boldsymbol{A}^{\delta}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) = {}&\left(\frac{(k\delta)^2}{4\pi}\,\boldsymbol{\beta}\times\left\{\boldsymbol{\beta}\times\int_{\Gamma}\left[\boldsymbol{\nu}_{\boldsymbol{\xi}}\times\left(\boldsymbol{h}^{\delta}(\boldsymbol{\xi})+\boldsymbol{h}_I^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{\xi})\right)\right]\,\mathrm{d}\gamma_{\boldsymbol{\xi}}\right\}\right)\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{s}}\\&+O(\delta^3).\end{aligned} \tag{3.7}$$

Comparing (3.6) with the term between the parentheses in the above expression, we see that this term is nothing but the electric far field associated to the solution $(\boldsymbol{e}^{\delta},\boldsymbol{h}^{\delta})$ of the low frequency scattering problem (3.2). Consequently, this term can be expressed using the polarizability tensors (see equation (5.158) in [12])

$$\begin{aligned}&\frac{(k\delta)^2}{4\pi}\,\boldsymbol{\beta}\times\left\{\boldsymbol{\beta}\times\int_{\Gamma}\left[\boldsymbol{\nu}_{\boldsymbol{\xi}}\times\left(\boldsymbol{h}^{\delta}(\boldsymbol{\xi})+\boldsymbol{h}_I^{\boldsymbol{\alpha},\boldsymbol{f}}(\boldsymbol{\xi})\right)\right]\,\mathrm{d}\gamma_{\boldsymbol{\xi}}\right\}\\&\qquad=\frac{(ik\delta)^3}{4\pi}\boldsymbol{\beta}\times\left[\boldsymbol{\beta}\times(\boldsymbol{\mathcal{P}}\boldsymbol{f}(\boldsymbol{\alpha}))-\boldsymbol{\mathcal{M}}(\boldsymbol{\alpha}\times\boldsymbol{f}(\boldsymbol{\alpha}))\right]\mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot\boldsymbol{s}}+O\left(\delta^4\right),\end{aligned}$$

where we have used the fact that the incident electromagnetic field $(\boldsymbol{e}_I,\boldsymbol{h}_I)$ converges to the constants $(\boldsymbol{E}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}),\boldsymbol{H}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{s}))=(\boldsymbol{f}(\boldsymbol{\alpha}),\boldsymbol{\alpha}\times\boldsymbol{f}(\boldsymbol{\alpha}))\,\mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot\boldsymbol{s}}$ as δ tends to 0. Plugging the last relation into (3.7) yields (3.5). □

3.2. Multiply connected scatterer. We consider now the case where the scatterer has M connected components:

$$\mathcal{O}^{\delta}=\bigcup_{p=1}^{M}\mathcal{O}_p^{\delta},$$

where each component $\mathcal{O}_p^{\delta}$ is obtained from a reference domain $\mathcal{O}_p$ by a dilation and a translation,

$$\mathcal{O}_p^{\delta}=\{\boldsymbol{x}=\boldsymbol{s}_p+\delta\boldsymbol{\xi}\ ;\ \boldsymbol{\xi}\in\mathcal{O}_p\}.$$

Finally, we denote once again by $\Omega^{\delta}=\mathbb{R}^3\setminus\overline{\mathcal{O}^{\delta}}$ the exterior domain and by $\Gamma^{\delta}=\bigcup_{p=1}^{M}\Gamma_p^{\delta}$ its boundary.

In order to study the asymptotics $\delta\to 0$, we seek an integral representation of the solution $(\boldsymbol{E}^{\delta},\boldsymbol{H}^{\delta})$ of (3.1) in the form

$$\left\{\begin{aligned}\boldsymbol{E}^{\delta}(\boldsymbol{y})&=\delta\,\mathrm{curl}\,\mathrm{curl}\int_{\Gamma^{\delta}}G_k(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{J}^{\delta}(\boldsymbol{x})\,\mathrm{d}\gamma_{\boldsymbol{x}},\\ \boldsymbol{H}^{\delta}(\boldsymbol{y})&=-\delta\,(ik)\,\mathrm{curl}\int_{\Gamma^{\delta}}G_k(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{J}^{\delta}(\boldsymbol{x})\,\mathrm{d}\gamma_{\boldsymbol{x}},\end{aligned}\right.\qquad \boldsymbol{y}\in\Omega^{\delta}, \tag{3.8}$$

where $\boldsymbol{J}^{\delta}$ is the (unknown) electric surface current and

$$G_k(\boldsymbol{x},\boldsymbol{y})=\frac{\mathrm{e}^{ik|\boldsymbol{x}-\boldsymbol{y}|}}{4\pi|\boldsymbol{x}-\boldsymbol{y}|}$$

denotes the Green function of $-\Delta-k^2$ in $\mathbb{R}^3$.

Using the identity curl curl $= \nabla \operatorname{div} - \Delta$ and the fact that for $\boldsymbol{x} \neq \boldsymbol{y}$ we have $\Delta G_k(\boldsymbol{x},\boldsymbol{y}) = -k^2\, G_k(\boldsymbol{x},\boldsymbol{y})$, one can show that the electric field can also be written in the form [10, p. 64]

$$\boldsymbol{E}^\delta(\boldsymbol{y}) = \delta \left(k^2 \int_{\Gamma^\delta} G_k(\boldsymbol{x},\boldsymbol{y})\, \boldsymbol{J}^\delta(\boldsymbol{x})\, \mathrm{d}\gamma_{\boldsymbol{x}} + \nabla \int_{\Gamma^\delta} G_k(\boldsymbol{x},\boldsymbol{y}) \operatorname{div}_{\Gamma^\delta} \boldsymbol{J}^\delta(\boldsymbol{x})\, \mathrm{d}\gamma_{\boldsymbol{x}} \right), \tag{3.9}$$

where $\operatorname{div}_{\Gamma^\delta}$ denotes the surface divergence operator on Γ^δ.

The unknown current $\boldsymbol{J}^\delta = (\boldsymbol{J}_1^\delta, \dots, \boldsymbol{J}_M^\delta)$ is uniquely determined by writing the perfectly conducting boundary condition on each scatterer:

$$(\boldsymbol{E}^\delta \times \boldsymbol{\nu})_{|\Gamma_p^\delta} = -(\boldsymbol{E}_I^{\alpha,\boldsymbol{f}} \times \boldsymbol{\nu})_{|\Gamma_p^\delta} \qquad \forall\, p = 1, \dots, M. \tag{3.10}$$

It is well known (see, for instance, [27, Theorem 5.5.1]) that the trace of a potential of the form (3.9) is given by

$$(\boldsymbol{E}^\delta \times \boldsymbol{\nu})_{|\Gamma_p^\delta} = \sum_{q=1}^{M} \delta \left(k^2 S_{pq}^{k,\delta} + T_{pq}^{k,\delta} \right) \boldsymbol{J}_q^\delta,$$

where the integral operators $S_{pq}^{k,\delta} : TH^s(\Gamma_q) \to TH^{s+1}(\Gamma_p)$ and $T_{pq}^{k,\delta} : TH^s(\Gamma_q) \to TH^{s-1}(\Gamma_p)$ ($TH^s(\Gamma_q)$ denotes the Sobolev space of tangent vector fields [27]) are defined for $\boldsymbol{y} \in \Gamma_p$ by

$$\left\{ \begin{array}{lcl} \left(S_{pq}^{k,\delta}\, \boldsymbol{J}_q^\delta\right)(\boldsymbol{y}) & = & \displaystyle\int_{\Gamma_q^\delta} G_k(\boldsymbol{x},\boldsymbol{y}) \left(\boldsymbol{J}_q^\delta(\boldsymbol{x}) \times \boldsymbol{\nu}_{\boldsymbol{y}}\right) \mathrm{d}\gamma_{\boldsymbol{x}}, \\ \left(T_{pq}^{k,\delta}\, \boldsymbol{J}_q^\delta\right)(\boldsymbol{y}) & = & \left(\nabla_{\boldsymbol{y}} \displaystyle\int_{\Gamma_q^\delta} G_k(\boldsymbol{x},\boldsymbol{y}) \operatorname{div}_{\Gamma_q^\delta} \boldsymbol{J}_q^\delta(\boldsymbol{x})\, \mathrm{d}\gamma_{\boldsymbol{x}}\right)_{|\Gamma_p^\delta} \times \boldsymbol{\nu}_{\boldsymbol{y}}. \end{array} \right.$$

For $q \neq p$, the kernels of the above integral operators are infinitely differentiable. The operator S_{pp}^δ is the classical single layer potential and has a singular but integrable kernel. The operator T_{pp}^δ can also be written using a formula involving only integrable kernels (see [27, p. 242]):

$$\begin{aligned} \left(T_{pp}^{k,\delta}\, \boldsymbol{J}_p^\delta\right)(\boldsymbol{y}) = \int_{\Gamma_p^\delta} \Big[& (\nabla_{\boldsymbol{y}} G_k(\boldsymbol{x},\boldsymbol{y}) \times (\boldsymbol{\nu}_{\boldsymbol{y}} - \boldsymbol{\nu}_{\boldsymbol{x}})) \operatorname{div}_{\Gamma_p^\delta} \boldsymbol{J}_p^\delta(\boldsymbol{x}) \\ & - G_k(\boldsymbol{x},\boldsymbol{y}) \operatorname{curl}_{\Gamma_p^\delta} \operatorname{div}_{\Gamma_p^\delta} \boldsymbol{J}_p^\delta(\boldsymbol{x}) \Big]\, \mathrm{d}\gamma_{\boldsymbol{x}}. \end{aligned}$$

In the above relation, $\operatorname{div}_{\Gamma_p^\delta}$ and $\operatorname{curl}_{\Gamma_p^\delta}$ denote, respectively, the surface divergence operator and tangential rotational operator on Γ_p^δ. Then, the integral equation (3.10) reads

$$\sum_{q=1}^{M} \delta \left(k^2 S_{pq}^{k,\delta} + T_{pq}^{k,\delta} \right) \boldsymbol{J}_q^\delta = -(\boldsymbol{E}^{inc} \times \boldsymbol{\nu})_{|\Gamma_p^\delta} \qquad \forall\, p = 1, \dots, M. \tag{3.11}$$

In order to work in a functional framework independent of δ, we introduce the new variables

$$\left\{ \begin{array}{l} \boldsymbol{\xi} = \dfrac{\boldsymbol{x} - \boldsymbol{s}_q}{\delta} \in \mathcal{O}_q, \\ \boldsymbol{\eta} = \dfrac{\boldsymbol{y} - \boldsymbol{s}_p}{\delta} \in \mathcal{O}_p \end{array} \right.$$

and the scaled fields

$$\begin{cases} \boldsymbol{j}_q^{\delta}(\boldsymbol{\xi}) = \boldsymbol{J}_q^{\delta}(\boldsymbol{x}), \\ \mathcal{G}_{pq}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta}) = G_k(\boldsymbol{x},\boldsymbol{y}). \end{cases}$$

With the above notation, we have

$$\left(S_{pq}^{k,\delta}\,\boldsymbol{J}_q^{\delta}\right)(\boldsymbol{y}) := \delta^2\left(\mathcal{S}_{pq}^{k,\delta}\,\boldsymbol{j}_q^{\delta}\right)(\boldsymbol{\eta}) = \delta^2\int_{\Gamma_q}\mathcal{G}_{pq}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\,\left(\boldsymbol{j}_q^{\delta}(\boldsymbol{\xi})\times\boldsymbol{\nu}_{\boldsymbol{\eta}}\right)\,\mathrm{d}\gamma_{\boldsymbol{\xi}}$$

and

$$\begin{aligned} \left(T_{pq}^{k,\delta}\,\boldsymbol{J}_p^{\delta}\right)(\boldsymbol{y}) &= \begin{cases} \left(\nabla_{\boldsymbol{\eta}}\displaystyle\int_{\Gamma_q}\mathcal{G}_{pq}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\,\mathrm{div}_{\Gamma_q}\boldsymbol{j}_q^{\delta}(\boldsymbol{\xi})\,\mathrm{d}\gamma_{\boldsymbol{\xi}}\right)_{|\Gamma_p}\times\boldsymbol{\nu}_{\boldsymbol{\eta}} & \text{for } q\neq p, \\ \displaystyle\int_{\Gamma_q}\Big[\left(\nabla_{\boldsymbol{\eta}}\mathcal{G}_{pp}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\times(\boldsymbol{\nu}_{\boldsymbol{\eta}}-\boldsymbol{\nu}_{\boldsymbol{\xi}})\right)\,\mathrm{div}_{\Gamma_q}\,\boldsymbol{j}_p^{\delta}(\boldsymbol{\xi}) & \\ \qquad\qquad -\,\mathcal{G}_{pp}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\,\mathrm{curl}_{\Gamma_p}\mathrm{div}_{\Gamma_p}\,\boldsymbol{j}_p^{\delta}(\boldsymbol{\xi})\Big]\,\mathrm{d}\gamma_{\boldsymbol{\xi}} & \text{for } q=p \end{cases} \\ &:= \left(\mathcal{T}_{pq}^{k,\delta}\,\boldsymbol{j}_q^{\delta}\right)(\boldsymbol{\eta}). \end{aligned}$$

Consequently, (3.11) can be written

$$\mathcal{B}_{pp}^{k,\delta}\,\boldsymbol{j}_p^{\delta} + \sum_{q\neq p}\mathcal{B}_{pq}^{k,\delta}\,\boldsymbol{j}_q^{\delta} = -(\boldsymbol{e}_I^{\alpha,\boldsymbol{f}}\times\boldsymbol{\nu})_{|\Gamma_p} \qquad \forall\, p=1,\dots,M, \tag{3.12}$$

with

$$\mathcal{B}_{pq}^{k,\delta} = (k\delta)^2\;\delta\,\mathcal{S}_{pq}^{k,\delta} + \delta\,\mathcal{T}_{pq}^{k,\delta} \tag{3.13}$$

and

$$\boldsymbol{e}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{\eta}) = \boldsymbol{E}_I^{\alpha,\boldsymbol{f}}(\boldsymbol{y}).$$

Let us consider first the diagonal terms in (3.12) by investigating the behavior of the kernels involved in the expression of $\mathcal{B}_{pp}^{k,\delta}$ as $\delta\to 0$. Since

$$\begin{cases} \mathcal{G}_{pp}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta}) = \dfrac{1}{\delta}G_{k\delta}(\boldsymbol{\xi},\boldsymbol{\eta}), \\ \nabla_{\boldsymbol{\eta}}\mathcal{G}_{pp}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta}) = \dfrac{1}{\delta}\nabla_{\boldsymbol{\eta}}G_{k\delta}(\boldsymbol{\xi},\boldsymbol{\eta}), \end{cases}$$

we see that

$$\mathcal{B}_{pp}^{k,\delta} = (k\delta)^2\,\widetilde{\mathcal{S}}_{pp}^{k\delta} + \widetilde{\mathcal{T}}_{pp}^{k\delta} := \widetilde{\mathcal{B}}_{pp}^{k\delta},$$

where

$$\begin{cases} \left(\widetilde{\mathcal{S}}_{pp}^{k\delta}\,\boldsymbol{j}_p\right)(\boldsymbol{y}) = \displaystyle\int_{\Gamma_q}G_{k\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\,\left(\boldsymbol{j}_p(\boldsymbol{\xi})\times\boldsymbol{\nu}_{\boldsymbol{\eta}}\right)\,\mathrm{d}\gamma_{\boldsymbol{\xi}}, \\ \left(\widetilde{\mathcal{T}}_{pp}^{k\delta}\,\boldsymbol{j}_q\right)(\boldsymbol{y}) = \displaystyle\int_{\Gamma_q}\Big[(\nabla_{\boldsymbol{\eta}}G_{k\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\times(\boldsymbol{\nu}_{\boldsymbol{\eta}}-\boldsymbol{\nu}_{\boldsymbol{\xi}}))\,\mathrm{div}_{\Gamma_p}\,\boldsymbol{j}_p(\boldsymbol{\xi}) \\ \qquad\qquad\qquad -\,G_{k\delta}(\boldsymbol{\xi},\boldsymbol{\eta})\,\mathrm{curl}_{\Gamma_p}\mathrm{div}_{\Gamma_p}\,\boldsymbol{j}_p(\boldsymbol{\xi})\Big]\,\mathrm{d}\gamma_{\boldsymbol{\xi}}. \end{cases}$$

The crucial point here is that $\widetilde{\mathcal{B}}_{pp}^{k\delta}$ is exactly the operator involved in the integral equation formulation of the simple scattering problem associated with the reference scatterer $\mathcal{O}_p$ at low frequency $k\delta \to 0$. Moreover, since the zero frequency limit exists, $\widetilde{\mathcal{B}}_{pp}^{k\delta} = \mathcal{B}_{pp}^{k,\delta}$ admits a limit $\mathcal{B}_{pp}^{0}$.

Let us consider now the off diagonal terms $\mathcal{B}_{pq}^{k,\delta}$, $q \neq p$. Denote by

$$d = \min_{1 \le p < q \le N} |\boldsymbol{s}_p - \boldsymbol{s}_q| \tag{3.14}$$

the minimal distance between the centers of the obstacles. Using the relation

$$|s_p - s_q + \delta(\boldsymbol{\xi} - \boldsymbol{\eta})| = |s_p - s_q| \left(1 + O\left(\frac{\delta}{d}\right)\right),$$

one can easily check that

$$\left\{\begin{aligned} &\mathcal{G}_{pq}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta}) = G_k(\boldsymbol{s}_q, \boldsymbol{s}_p)\left[1 + O(k\delta) + O\left(\frac{\delta}{d}\right)\right], \\ &\nabla_{\boldsymbol{\eta}}\, \mathcal{G}_{pq}^{k,\delta}(\boldsymbol{\xi},\boldsymbol{\eta}) = G_k(\boldsymbol{s}_q, \boldsymbol{s}_p)\left[O(k\delta) + O\left(\frac{\delta}{d}\right)\right] \end{aligned}\right. \qquad \forall q \neq p.$$

Inserting the above asymptotics into (3.13) shows that

$$\mathcal{B}_{pq}^{k,\delta} = O\left(\frac{\delta}{d}\right)\left[O(k\delta) + O\left(\frac{\delta}{d}\right)\right] \qquad \text{for } q \neq p.$$

Summing up, the behavior of the solution $(\boldsymbol{E}^{\delta}, \boldsymbol{H}^{\delta})$ of (3.1) for small scatterers (namely, for $k\delta \to 0$ and $\delta/d \to 0$) is given by the low frequency limit of the simple scattering problem. Therefore, the multiple scattering effects can be neglected when $k\delta \to 0$ and $\delta/d \to 0$, and the electric far field can be obtained simply by superposition of the far fields given in Corollary 3.2. We have thus proved the following result.

THEOREM 3.3. *Assume that the scatterer has M connected components*

$$\mathcal{O}^{\delta} = \bigcup_{p=1}^{M} \mathcal{O}_p^{\delta},$$

where each component $\mathcal{O}_p^{\delta}$ is obtained from a reference scatterer $\mathcal{O}_p$ (centered at the origin) of smooth boundary Γ_p by a dilation of ratio δ centered at a given point $\boldsymbol{s}_p \in \mathbb{R}^3$:

$$\mathcal{O}_p^{\delta} = \{\boldsymbol{x} = \boldsymbol{s}_p + \delta\boldsymbol{\xi}\ ;\ \boldsymbol{\xi} \in \mathcal{O}_p\}.$$

For all $p = 1, \ldots, M$, let $\boldsymbol{\Phi}_p$ and $\boldsymbol{\Psi}_p$ be the vector potentials defined by

$$\left\{\begin{array}{ll} \Delta\boldsymbol{\Phi}_p = 0 & (\mathbb{R}^3 \setminus \overline{\mathcal{O}_p}), \\ \boldsymbol{\Phi}_p = \boldsymbol{x} + \boldsymbol{c}_p & (\Gamma_p), \\ \boldsymbol{\Phi}_p = O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), & |\boldsymbol{x}| \to \infty, \end{array}\right. \tag{3.15}$$

and

$$\left\{\begin{array}{ll}\Delta\boldsymbol{\Psi}_p = 0 & (\mathbb{R}^3\setminus\overline{\mathcal{O}_p}),\\ \dfrac{\partial\boldsymbol{\Psi}_p}{\partial\boldsymbol{\nu}} = \boldsymbol{\nu} & (\Gamma_p),\\ \boldsymbol{\Psi}_p = O\left(\dfrac{1}{|\boldsymbol{x}|^2}\right), & |\boldsymbol{x}|\to\infty,\end{array}\right. \tag{3.16}$$

where the constant vector $\boldsymbol{c}_p \in \mathbb{R}^3$ *is chosen such that* $\int_{\Gamma_p}\frac{\partial\boldsymbol{\Phi}_p}{\partial\boldsymbol{\nu}} = 0$.

Let $\boldsymbol{\mathcal{P}}_p$ *and* $\boldsymbol{\mathcal{M}}_p$ *be, respectively, the electric polarizability and magnetic polarizability tensors of the reference scatterer* $\mathcal{O}_p$ *(*I *denotes the identity):*

$$\left\{\begin{array}{l}\boldsymbol{\mathcal{P}}_p = |\mathcal{O}_p|\, I - \displaystyle\int_{\Gamma_p}\boldsymbol{x}\left(\frac{\partial\boldsymbol{\Phi}_p}{\partial\boldsymbol{\nu}}\right)^T \mathrm{d}\gamma_{\boldsymbol{x}},\\ \boldsymbol{\mathcal{M}}_p = |\mathcal{O}_p|\, I - \displaystyle\int_{\Gamma_p}\boldsymbol{\nu}\boldsymbol{\Psi}_p^T\,\mathrm{d}\gamma_{\boldsymbol{x}}.\end{array}\right. \tag{3.17}$$

Finally, let $(\boldsymbol{E}^\delta, \boldsymbol{H}^\delta)$ *be the solution of the scattering problem* (3.1) *and* $\boldsymbol{A}^\delta(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))$ *the far field of* $\boldsymbol{E}^\delta$*:*

$$\boldsymbol{E}^\delta(\boldsymbol{\beta}|\boldsymbol{x}|) = \boldsymbol{A}^\delta(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))\frac{\mathrm{e}^{ik|\boldsymbol{x}|}}{ik|\boldsymbol{x}|} + O\left(\frac{1}{|\boldsymbol{x}|^2}\right).$$

Then, as $\delta\to 0$*, we have*

$$\frac{4\pi}{(ik\delta)^3}\boldsymbol{A}^\delta(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) \longrightarrow \boldsymbol{A}^0(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})), \tag{3.18}$$

where

$$\boldsymbol{A}^0(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha})) = \sum_{p=1}^{M}\boldsymbol{\beta}\times\Big[\boldsymbol{\beta}\times(\boldsymbol{\mathcal{P}}_p\boldsymbol{f}(\boldsymbol{\alpha})) - \boldsymbol{\mathcal{M}}_p(\boldsymbol{\alpha}\times\boldsymbol{f}(\boldsymbol{\alpha}))\Big]\mathrm{e}^{ik\,(\boldsymbol{\alpha}-\boldsymbol{\beta})\cdot\boldsymbol{s}_p}. \tag{3.19}$$

The convergence (3.18) *holds uniformly for all* $\alpha,\beta\in S^2$ *and for all wavenumbers* k *and minimal distances* d *(defined by* (3.14)*) satisfying* $k\delta\to 0$ *and* $\delta/d\to 0$.

4. Selective focusing using time reversal. From now on, we assume that $k\delta\to 0$ and $\delta/d\to 0$. According to Theorem 3.3, the eigenfunctions of the far field operator $\mathbf{F}^\delta$ can be approximated by those of the operator $\mathbf{F}^0 : L^2_t(S^2)\to L^2_t(S^2)$ defined by

$$(\mathbf{F}^0\boldsymbol{f})(\boldsymbol{\beta}) = \int_{S^2}\boldsymbol{A}^0(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))\,\mathrm{d}\boldsymbol{\alpha} \qquad \forall\,\boldsymbol{f}\in L^2_t(S^2). \tag{4.1}$$

Substituting the expression (3.19) of $\boldsymbol{A}^0(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{f}(\boldsymbol{\alpha}))$ into (4.1), we obtain that

$$(\mathbf{F}^0\boldsymbol{f})(\boldsymbol{\beta}) = \sum_{p=1}^{M}\boldsymbol{\beta}\times\Big[\boldsymbol{\beta}\times\Big(\boldsymbol{\mathcal{P}}_p\boldsymbol{E}_I^{\boldsymbol{f}}(\boldsymbol{s}_p)\Big) - \boldsymbol{\mathcal{M}}_p\boldsymbol{H}_I^{\boldsymbol{f}}(\boldsymbol{s}_p)\Big]\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{s}_p}, \tag{4.2}$$

where $(\boldsymbol{E}_I^{\boldsymbol{f}}, \boldsymbol{H}_I^{\boldsymbol{f}})$ denote the electromagnetic Herglotz wave associated to $\boldsymbol{f}$ defined by (2.5). Finally, let us notice that

$$(\mathbf{F}^0\boldsymbol{f})(\boldsymbol{\beta}) = -\sum_{p=1}^{M}\Big[\Delta(\boldsymbol{\beta})\boldsymbol{\mathcal{P}}_p\boldsymbol{E}_I^{\boldsymbol{f}}(\boldsymbol{s}_p) + \boldsymbol{\beta}\times(\boldsymbol{\mathcal{M}}_p\boldsymbol{H}_I^{\boldsymbol{f}}(\boldsymbol{s}_p))\Big]\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{s}_p}, \tag{4.3}$$

where for every

$$\boldsymbol{\alpha} = \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix} \in S^2$$

we have set

$$\Delta(\boldsymbol{\alpha}) = \mathbf{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T. \tag{4.4}$$

REMARK 4.1. *Note that formula* (4.2) *shows that the electric far field radiated by the scatterers as their size tends to* 0 *corresponds to the superposition of* M *(uncoupled) electric and magnetic dipoles located at the points* $\boldsymbol{s}_p$ *and associated with the electric and magnetic moments* $\boldsymbol{p}_p = \boldsymbol{\mathcal{P}}_p\, \boldsymbol{E}_I^{\boldsymbol{f}}(\boldsymbol{s}_p)$ *and* $\boldsymbol{m}_p = \boldsymbol{\mathcal{M}}_p\, \boldsymbol{H}_I^{\boldsymbol{f}}(\boldsymbol{s}_p)$.

REMARK 4.2. *Formula* (4.3) *shows that* $\mathbf{F}^0$ *has at most* $6M$ *nonzero eigenvalues, since its range satisfies* $\operatorname{Ran}\mathbf{F}^0 \subset \bigoplus_{1\le p\le M} \left\{ \left(\Delta(\boldsymbol{\beta})\operatorname{Ran}\boldsymbol{\mathcal{P}}_p\right) \oplus \left(\boldsymbol{\beta}\times\operatorname{Ran}\boldsymbol{\mathcal{M}}_p\right)\right\}$.

The aim of this section is twofold: first, to compute approximate eigenfunctions of $\mathbf{F}^0$, and then to prove that these eigenfunctions selectively focus on the scatterers. As we will see, this can be achieved provided the following two assumptions are satisfied:

1. The polarizability tensors $\boldsymbol{\mathcal{P}}_p$ and $\boldsymbol{\mathcal{M}}_p$ are diagonal (in the same basis). This is particularly true for axially symmetric scatterers (see [12, p. 167]).
2. The scatterers are distant enough (well-separated scatterers). More precisely, we assume that $kd \to \infty$, where $d = \min_{1\le p<q\le N} |\boldsymbol{s}_p - \boldsymbol{s}_q|$ is the minimal distance between the obstacles.

From now on, we will assume that these two conditions are satisfied.

THEOREM 4.3. *For* $p \in \{1,\dots,M\}$, *let* $(\boldsymbol{e}_{p,1}, \boldsymbol{e}_{p,2}, \boldsymbol{e}_{p,3})$ *be an orthonormal basis of* $\mathbb{R}^3$ *such that the polarizability tensors* $\boldsymbol{\mathcal{P}}_p, \boldsymbol{\mathcal{M}}_p$ *of the reference scatterer* $\mathcal{O}_p$ *are diagonal:*

$$\boldsymbol{\mathcal{P}}_p = \begin{bmatrix} \lambda_{p,1} & 0 & 0 \\ 0 & \lambda_{p,2} & 0 \\ 0 & 0 & \lambda_{p,3} \end{bmatrix}, \qquad \boldsymbol{\mathcal{M}}_p = \begin{bmatrix} \lambda'_{p,1} & 0 & 0 \\ 0 & \lambda'_{p,2} & 0 \\ 0 & 0 & \lambda'_{p,3} \end{bmatrix}. \tag{4.5}$$

Given $\ell \in \{1,2,3\}$, *define the following elements of* $L^2_t(S^2)$ *(recall that* $\Delta(\boldsymbol{\alpha})$ *is defined by* (4.4)*):*

$$\left\{ \begin{aligned} &\boldsymbol{f}_{p,\ell}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}\times(\boldsymbol{\alpha}\times\boldsymbol{e}_{p,\ell})\,\mathrm{e}^{-ik\,\boldsymbol{\alpha}\cdot\boldsymbol{s}_p} = -\Delta(\boldsymbol{\alpha})\boldsymbol{e}_{p,\ell}\,\mathrm{e}^{-ik\,\boldsymbol{\alpha}\cdot\boldsymbol{s}_p}, \\ &\boldsymbol{g}_{p,\ell}(\boldsymbol{\alpha}) = (\boldsymbol{\alpha}\times\boldsymbol{e}_{p,\ell})\,\mathrm{e}^{-ik\,\boldsymbol{\alpha}\cdot\boldsymbol{s}_p}, \end{aligned} \right. \qquad \boldsymbol{\alpha}\in S^2. \tag{4.6}$$

Then, the family of functions $\{\boldsymbol{f}_{p,\ell}, \boldsymbol{g}_{p,\ell};\ 1\le\ell\le 3,\ 1\le p\le M\}$ *is linearly independent in* $L^2_t(S^2)$. *Moreover, the functions* $\boldsymbol{f}_{p,\ell}$ *and* $\boldsymbol{g}_{p,\ell}$ *constitute approximate eigenfunctions of the limit far field operator* $\mathbf{F}^0$ *defined by* (3.19)–(4.1) *as* $kd\to\infty$*:*

$$\left\{ \begin{aligned} &\mathbf{F}^0\boldsymbol{f}_{p,\ell} = -\frac{8\pi}{3}\lambda_{p,\ell}\ \boldsymbol{f}_{p,\ell} + O\left((kd)^{-N}\right), \\ &\mathbf{F}^0\boldsymbol{g}_{p,\ell} = -\frac{8\pi}{3}\lambda'_{p,\ell}\ \boldsymbol{g}_{p,\ell} + O\left((kd)^{-N}\right) \end{aligned} \right. \qquad \forall N\in\mathbb{N}. \tag{4.7}$$

Proof. To see that the functions $\boldsymbol{f}_{p,\ell}$ and $\boldsymbol{g}_{p,\ell}$, for $\ell = 1,2,3$ and $p = 1,\dots,M$ are linearly independent, it suffices to note that these functions are exactly the far field patterns of electric and magnetic dipoles located at the points $\boldsymbol{s}_p$ and associated with electric or magnetic dipole moment $\boldsymbol{e}_{p,\ell}$. Consequently, by uniqueness of the far field pattern (which follows from Rellich's lemma; see [11]), the condition

$$\sum_{p=1}^{M}\sum_{\ell=1}^{3}(z_{p,\ell}\,\boldsymbol{f}_{p,\ell}+z'_{p,\ell}\,\boldsymbol{g}_{p,\ell})=0,\qquad z_{p,\ell},\,z'_{p,\ell}\in\mathbb{C},$$

implies that $z_{p,\ell}=z'_{p,\ell}=0$ for all $p=1,\dots,M$ and $\ell=1,2,3$.

Fix now $q\in\{1,\dots,M\}$ and $\ell\in\{1,2,3\}$, and let us compute $\mathbf{F}^0\boldsymbol{f}_{q,\ell}$. We have

$$\begin{cases} \boldsymbol{E}_I^{\boldsymbol{f}_{q,\ell}}(\boldsymbol{s}_p)=-\left(\displaystyle\int_{S^2}\Delta(\boldsymbol{\alpha})\,\mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot(\boldsymbol{s}_p-\boldsymbol{s}_q)}\,\mathrm{d}\boldsymbol{\alpha}\right)\boldsymbol{e}_{q,\ell}:=\boldsymbol{D}_{pq}\,\boldsymbol{e}_{q,\ell},\\ \boldsymbol{H}_I^{\boldsymbol{f}_{q,\ell}}(\boldsymbol{s}_p)=\displaystyle\int_{S^2}\boldsymbol{\alpha}\times[\boldsymbol{\alpha}\times(\boldsymbol{\alpha}\times\boldsymbol{e}_{q,\ell})]\,\mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot(\boldsymbol{s}_p-\boldsymbol{s}_q)}\,\mathrm{d}\boldsymbol{\alpha}:=\boldsymbol{D}'_{pq}\,\boldsymbol{e}_{q,\ell}. \end{cases}$$

A straightforward computation shows that

$$\boldsymbol{D}_{qq}=-\int_{S^2}\Delta(\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\alpha}=-\frac{8\pi}{3}\mathbf{I},$$

while by symmetry

$$\boldsymbol{D}'_{qq}=\int_{S^2}\boldsymbol{\alpha}\times\big[\boldsymbol{\alpha}\times(\boldsymbol{\alpha}\times\boldsymbol{e}_{q,\ell})\big]\,\mathrm{d}\boldsymbol{\alpha}=0.$$

On the other hand, let us note that the elements of the 3×3 matrices $\boldsymbol{D}_{pq}$ and $\boldsymbol{D}'_{pq}$ for $p\neq q$ are oscillatory integrals of the form $\int_{S^2}\psi(\boldsymbol{\alpha})\,\mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot(\boldsymbol{s}_p-\boldsymbol{s}_q)}\,\mathrm{d}\boldsymbol{\alpha}$, where ψ is a smooth function. It follows then from the stationary phase theorem (see, for instance, [33, Chap. VIII]) that

$$\boldsymbol{D}_{pq}=\boldsymbol{D}'_{pq}=O\left((kd)^{-N}\right)\qquad\forall\,p\neq q,\ \forall N\in\mathbb{N}.$$

Consequently, formula (4.3) simplifies to

$$\begin{aligned}(\mathbf{F}^0\boldsymbol{f}_{q,\ell})(\boldsymbol{\beta}) &= -\frac{8\pi}{3}\Delta(\boldsymbol{\beta})\boldsymbol{\mathcal{P}}_q\boldsymbol{e}_{q,\ell}\,\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{s}_q}+O\left((kd)^{-N}\right),\\ &= -\frac{8\pi}{3}\lambda_{q,\ell}\,\Delta(\boldsymbol{\beta})\boldsymbol{e}_{q,\ell}\,\mathrm{e}^{-ik\,\boldsymbol{\beta}\cdot\boldsymbol{s}_q}+O\left((kd)^{-N}\right),\end{aligned}$$

which proves the first relation of (4.7). The second relation of (4.7) follows using the same arguments, since

$$\begin{cases}\boldsymbol{E}_I^{\boldsymbol{g}_{q,\ell}}(\boldsymbol{s}_p)=-\boldsymbol{H}_I^{\boldsymbol{f}_{q,\ell}}(\boldsymbol{s}_p)=-\boldsymbol{D}'_{pq}\,\boldsymbol{e}_{q,\ell},\\ \boldsymbol{H}_I^{\boldsymbol{g}_{q,\ell}}(\boldsymbol{s}_p)=\boldsymbol{E}_I^{\boldsymbol{f}_{q,\ell}}(\boldsymbol{s}_p)=\boldsymbol{D}_{pq}\,\boldsymbol{e}_{q,\ell},\end{cases}$$

and the proof is thus complete. □

REMARK 4.4. *In the special case of scattering by small triaxial ellipsoids (see* [12, Chap. 8]*) with semiaxes $a_{p,1}>a_{p,2}>a_{p,3}$, the electric and magnetic polarizability tensors admit in the basis constituted by the axis of each ellipsoid the diagonal form* (4.5)*, with*

$$\begin{cases}\lambda_{p,\ell}=\dfrac{4\pi}{3I_{p,\ell}},\\ \lambda'_{p,\ell}=\dfrac{4\pi}{3}\dfrac{a_{p,1}a_{p,2}a_{p,3}}{1-a_{p,1}a_{p,2}a_{p,3}I_{p,\ell}},\end{cases}\qquad \ell=1,2,3,$$

with

$$I_{p,\ell} = \frac{2\pi}{3} \int_0^\infty \frac{\mathrm{d}x}{(x + a_{p,\ell}^2)\sqrt{x^2 + a_{p,1}^2}\sqrt{x^2 + a_{p,2}^2}\sqrt{x^2 + a_{p,3}^2}}.$$

In the special case of spheres of radii a_p, *we have* $\boldsymbol{\mathcal{P}}_p = 2\boldsymbol{\mathcal{M}}_p = 4\pi a_p^3\, I$.

The next result provides the expected selective focusing properties of the eigenfunctions of the far field operator $\mathbf{F}^0$ (and thus of time reversal operator $\mathbf{T}^0 = (\mathbf{F}^0)^*\mathbf{F}^0$).

Theorem 4.5. *For* $p \in \{1, \dots, M\}$, *the approximate eigenfunctions* $(\boldsymbol{f}_{p,\ell}, \boldsymbol{g}_{p,\ell})_{1\le\ell\le3}$ *defined by* (4.6) *generate electromagnetic Herglotz waves that focus selectively on the scatterer* p.

Proof. Plugging the expression (4.6) of $\boldsymbol{f}_{p,\ell}$ and $\boldsymbol{g}_{p,\ell}$ into (2.5), we obtain that

$$\begin{cases} \boldsymbol{E}_I^{\boldsymbol{f}_{p,\ell}}(\boldsymbol{x}) = \boldsymbol{H}_I^{\boldsymbol{g}_{p,\ell}}(\boldsymbol{x}) = \displaystyle\int_{S^2} (\boldsymbol{\alpha} \times (\boldsymbol{\alpha} \times \boldsymbol{e}_{p,\ell}))\, \mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot(\boldsymbol{x}-\boldsymbol{s}_p)}\, \mathrm{d}\boldsymbol{\alpha}, \\ \boldsymbol{H}_I^{\boldsymbol{f}_{p,\ell}}(\boldsymbol{x}) = -\boldsymbol{E}_I^{\boldsymbol{g}_{p,\ell}}(\boldsymbol{x}) = -\displaystyle\int_{S^2} (\boldsymbol{\alpha} \times \boldsymbol{e}_{p,\ell})\, \mathrm{e}^{ik\,\boldsymbol{\alpha}\cdot(\boldsymbol{x}-\boldsymbol{s}_p)}\, \mathrm{d}\boldsymbol{\alpha}. \end{cases}$$

The conclusion follows once again from the stationary phase theorem, since for $\boldsymbol{x} \neq \boldsymbol{s}_p$ the above integrals behave like $O\left((k|x - s_p|)^{-N}\right)$ for all $N \in \mathbb{N}$ and are independent of k for $\boldsymbol{x} = \boldsymbol{s}_p$. □

REFERENCES

[1] H. Ammari, E. Iakovleva, D. Lesselier, and G. Perrusson, *MUSIC-type electromagnetic imaging of a collection of small three-dimensional inclusions*, SIAM J. Sci. Comput., 29 (2007), pp. 674–709.

[2] H. Ammari and H. Kang, *Generalized polarization tensors, inverse conductivity problems, and dilute composite materials: A review*, in Inverse Problems, Multi-Scale Analysis and Effective Medium Theory, Contemp. Math. 408, AMS, Providence, RI, 2006, pp. 1–67.

[3] H. Ammari and H. Kang, *Polarization and Moment Tensors. With Applications to Inverse Problems and Effective Medium Theory*, Appl. Math. Sci. 162, Springer-Verlag, New York, 2007.

[4] H. Ammari and J.-C. Nédélec, *Low-frequency electromagnetic scattering*, SIAM J. Math. Anal., 31 (2000), pp. 836–861.

[5] F. Cakoni and D. Colton, *Combined far-field operators in electromagnetic inverse scattering theory*, Math. Methods Appl. Sci., 26 (2003), pp. 413–429.

[6] D. H. Chambers and J. G. Berryman, *Analysis of the time-reversal operator for a small spherical scatterer in an electromagnetic field*, IEEE Trans. Antennas and Propagation, 52 (2004), pp. 1729–1738.

[7] D. H. Chambers and J. G. Berryman, *Target characterization using decomposition of the time-reversal operator: Electromagnetic scattering from small ellipsoids*, Inverse Problems, 22 (2006), pp. 2145–2163.

[8] M. Cheney and G. Kristensson, *Optimal Electromagnetic Measurements*, Tech. Report TEAT-7091, Lund Institute of Technology (LUTEDX), Lund, Sweden, 2001.

[9] D. Colton and R. Kress, *Eigenvalues of the far field operator and inverse scattering theory*, SIAM J. Math. Anal., 26 (1995), pp. 601–615.

[10] D. L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Pure and Applied Mathematics (New York), John Wiley & Sons, New York, 1983.

[11] D. L. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[12] G. Dassios and R. Kleinman, *Low Frequency Scattering*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 2000.

[13] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.-L. Thomas, and F. Wu, *Time-reversed acoustics*, Rep. Progr. Phys., 63 (2000), pp. 1933–1995.

[14] M. Fink and C. Prada, *Eigenmodes of the time-reversal operator: A solution to selective focusing in multiple-target media*, Wave Motion, 20 (1994), pp. 151–163.

[15] M. Fink and C. Prada, *Acoustic time-reversal mirrors*, Inverse Problems, 17 (2001), pp. 1761–1773.

[16] T. Folégot, C. Prada, and M. Fink, *Resolution enhancement and separation of reverberation from target echo with the time reversal operator decomposition*, J. Acoust. Soc. Am., 113 (2003), pp. 3155–3160.

[17] J.-P. Fouque, J. Garnier, G. Papanicolaou, and K. Sølna, *Wave Propagation and Time Reversal in Randomly Layered Media*, Stoch. Model. Appl. Probab. 56, Springer-Verlag, New York, 2007.

[18] C. F. Gaumond, D. M. Fromm, J. F. Lingevitch, R. Menis, G. F. Edelmann, D. C. Calvo, and E. Kim, *Demonstration at sea of the decomposition-of-the-time-reversal-operator technique*, J. Acoust. Soc. Am., 119 (2006), pp. 976–990.

[19] C. Hazard and K. Ramdani, *Selective acoustic focusing using time-harmonic reversal mirrors*, SIAM J. Appl. Math., 64 (2004), pp. 1057–1076.

[20] A. Kirsch, *The factorization method for Maxwell's equations*, Inverse Problems, 20 (2004), pp. S117–S134.

[21] R. E. Kleinman and T. B. A. Senior, *Rayleigh scattering*, in Low and High Frequency Asymptotics, Mech. Math. Methods. Ser. Handbooks. Ser. III Acoust. Electromagnet. Elastic Wave Scatt. 2, North–Holland, Amsterdam, 1986, pp. 1–70.

[22] R. Kress, *On the limiting behaviour of solutions to boundary integral equations associated with time harmonic wave equations for small frequencies*, Math. Methods Appl. Sci., 1 (1979), pp. 89–100.

[23] G. Lerosey, J. de Rosny, A. Tourin, A. Derode, G. Montaldo, and M. Fink, *Time reversal of electromagnetic waves*, Phy. Rev. Lett., 92 (2004), 193904.

[24] D. Liu, G. Kang, L. Li, Y. Chen, S. Vasudevan, W. Joines, Q. Liu, J. Krolik, and L. Carin, *Electromagnetic time-reversal imaging of a target in a cluttered environment*, IEEE Trans. Antennas and Propagation, 53 (2005), pp. 3058–3066.

[25] J.-G. Minonzio, C. Prada, A. Aubry, and M. Fink, *Multiple scattering between two elastic cylinders and invariants of the time-reversal operator: Theory and experiment*, J. Acoust. Soc. Am., 120 (2006), pp. 875–883.

[26] N. Mordant, C. Prada, and M. Fink, *Highly resolved detection and selective focusing in a waveguide using the D.O.R.T. method*, J. Acoust. Soc. Am., 105 (1999), pp. 2634–2642.

[27] J.-C. Nédélec, *Acoustic and Electromagnetic Equations*, Appl. Math. Sci. 144, Springer-Verlag, New York, 2001.

[28] R. Picard, *On the low frequency asymptotics in electromagnetic theory*, J. Reine Angew. Math., 354 (1984), pp. 50–73.

[29] B. Pinçon and K. Ramdani, *Selective focusing on small scatterers in acoustic waveguides using time reversal mirrors*, Inverse Problems, 23 (2007), pp. 1–25.

[30] G. Pólya and G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*, Ann. of Math. Stud. 27, Princeton University Press, Princeton, NJ, 1951.

[31] C. Prada, E. Kerbrat, D. Cassereau, and M. Fink, *Time reversal techniques in ultrasonic nondestructive testing of scattering media*, Inverse Problems, 18 (2002), pp. 1761–1773.

[32] C. Prada, S. Manneville, D. Spoliansky, and M. Fink, *Decomposition of the time reversal operator: Detection and selective focusing on two scatterers*, J. Acoust. Soc. Am., 9 (1996), pp. 2067–2076.

[33] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton Math. Ser. 43, Princeton University Press, Princeton, NJ, 1993.

[34] H. Tortel, G. Micolau, and M. Saillard, *Decomposition of the time reversal operator for electromagnetic scattering*, J. Electromagn. Waves Appl., 13 (1999), pp. 687–719.

[35] P. Werner, *On the exterior boundary value problem of perfect reflection for stationary electromagnetic wave fields*, J. Math. Anal. Appl., 7 (1963), pp. 348–396.

[36] P. Werner, *On the behavior of stationary electromagnetic wave fields for small frequencies*, J. Math. Anal. Appl., 15 (1966), pp. 447–496.

SIAM J. APPL. MATH.
Vol. 69, No. 3, pp. 845–880

THE FRAMEWORK OF k-HARMONICALLY ANALYTIC FUNCTIONS FOR THREE-DIMENSIONAL STOKES FLOW PROBLEMS, PART I*

MICHAEL ZABARANKIN†

Abstract. The framework of generalized analytic functions arising from the related potentials (so-called k-harmonically analytic functions) has been developed in application to three-dimensional (3D) axially symmetric Stokes flow problems. Cauchy's integral formula for the class of k-harmonically analytic functions has been obtained, and series representations for k-harmonically analytic functions for the regions exterior to sphere and prolate and oblate spheroids have been derived. As the central result in the developed framework, a solution form representing the velocity field and pressure for 3D axially symmetric Stokes flows has been constructed in terms of two 0-harmonically analytic functions. It has also been shown that it uniquely determines an external velocity field vanishing at infinity. With the obtained solution form, the problem of 3D Stokes flows due to the axially symmetric translation of a solid body of revolution has been reduced to a boundary-value problem for two 0-harmonically analytic functions, and the resisting force exerted on the body has been expressed in terms of a 0-harmonically analytic function entering the solution form. For regions in which Laplace's equation admits separation of variables, the boundary-value problem can be solved analytically via representations of 0-harmonically analytic functions in corresponding curvilinear coordinates. This approach has been demonstrated for the axially symmetric translation of solid sphere and solid prolate and oblate spheroids. As the second approach, the boundary-value problem has been reduced to an integral equation based on Cauchy's integral formula for k-harmonically analytic functions. As an illustration, the integral equation has been solved for the axially symmetric translation of solid bispheroids and the solid torus of elliptical cross-section for various values of a geometrical parameter.

Key words. Stokes flows, generalized analytic functions, exact solution, generalized Cauchy's integral formula, integral equation

AMS subject classifications. 30E20, 35Q15, 35Q30, 76D07

DOI. 10.1137/080715913

1. Introduction.

This article is the first part of the developed two-part framework of k-harmonically analytic functions in application to three-dimensional (3D) Stokes flow problems. In this work, we obtain Cauchy's integral formula for k-harmonically analytic functions, construct series representations for k-harmonically analytic functions for the regions exterior to sphere and prolate and oblate spheroids, and develop the framework of 0-harmonically analytic functions for *axially symmetric* Stokes flow problems. Our second article, the sequel to this paper, develops the framework of k-harmonically analytic functions for *asymmetric* Stokes flow problems.

1.1. Generalized analytic functions.

Generalizations of the theory of functions of complex variables are represented mainly by the theories of generalized analytic functions (Vekua [24]), pseudoanalytic functions (Bers [5, 6]), and p-analytic functions (Položii [20]), which replace the classical Cauchy–Riemann system by certain systems of linear first-order partial differential equations (relating real and imaginary

*Received by the editors February 18, 2008; accepted for publication (in revised form) August 12, 2008; published electronically December 31, 2008. Research was partially supported by AFOSR grant FA9550-07-1-0401.

http://www.siam.org/journals/siap/69-3/71591.html

†Department of Mathematical Sciences, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030 (mzabaran@stevens.edu).

parts). Generalized analytic and pseudoanalytic[1] functions satisfy the same Bers–Vekua system [24, 5, 6], whereas a class of p-analytic functions $P(r,z) = \widehat{U}(r,z) + i\,\widehat{V}(r,z)$ is introduced via Položii's system, $\frac{\partial \widehat{U}}{\partial r} = \frac{1}{p}\,\frac{\partial \widehat{V}}{\partial z}$ and $\frac{\partial \widehat{U}}{\partial z} = -\frac{1}{p}\,\frac{\partial \widehat{V}}{\partial r}$, where $p = p(r,z) \geq 0$. The aforementioned theories furnish integral representations for generalized analytic functions via ordinary analytic functions and generalize a majority of fundamental classical results, including Cauchy's integral formula, Cauchy-type integrals, power series, etc.

Special cases of Bers–Vekua and Položii systems arise in various 3D problems of hydrodynamics, aerodynamics, elastic medium, electromagnetism, etc., and are associated with subclasses of generalized analytic functions, tailored for specific applications. For example, in application to mechanics of continua, Bers and Gelbart investigated a class of functions defined by the system of equations of "mixed type" [7, 8], [6, p. 30], whose generalized version was used to determine a class of Σ-monogenic functions [4]. A special case of Položii's system for $p = r^k$, known also as generalized Stokes–Beltrami equations, finds its application in the 3D axially symmetric theory of elasticity [20] and was also studied in the context of generalized axially symmetric potential theory [28]. The relationship between p-analytic functions and the Schrödinger equation was discussed in [13].

In [30], we introduced a special class of generalized analytic functions that arise from the fundamental relationship between a scalar field ϕ and vectorial field $\mathbf{\Lambda}$:

$$\operatorname{grad}\phi = -\operatorname{curl}\mathbf{\Lambda}, \qquad \operatorname{div}\mathbf{\Lambda} = 0, \tag{1}$$

which maintains that ϕ and $\mathbf{\Lambda}$ are related scalar and vectorial potentials, respectively. This relationship is encountered in various areas of applied mathematics, particularly in hydrodynamics, the theory of elasticity, electromagnetism, etc.; see [30].

Example 1 (Stokes flows). The behavior of steady flows of a viscous incompressible fluid under the assumption of a zero (low) Reynolds number (so-called *Stokes creeping flows*) is described by the Stokes equations

$$\mu\,\mathbf{\Delta u} = \operatorname{grad}\wp, \qquad \operatorname{div}\mathbf{u} = 0, \tag{2}$$

where $\mathbf{u}$ is the fluid velocity field, $\wp$ is the pressure in the fluid, μ is the shear viscosity, and $\mathbf{\Delta u} \equiv \operatorname{grad}(\operatorname{div}\mathbf{u}) - \operatorname{curl}\operatorname{curl}\mathbf{u}$; see [11, 14]. With $\operatorname{div}\mathbf{u} = 0$, the first equation in (2) can be rewritten as $\operatorname{grad}\wp = -\mu\operatorname{curl}(\operatorname{curl}\mathbf{u})$, whence it follows that the vorticity $\boldsymbol{\omega} = \operatorname{curl}\mathbf{u}$ and pressure $\wp$ are related by $\operatorname{grad}\wp = -\mu\operatorname{curl}\boldsymbol{\omega}$ with $\operatorname{div}\boldsymbol{\omega} = 0$, and thus, $\wp$ and $\mu\,\boldsymbol{\omega}$ are related potentials satisfying (1).

In the two-dimensional (2D) case in Cartesian coordinates, (1) reduces to the classical Cauchy–Riemann system for ordinary analytic functions, and in the 3D axially symmetric case in the cylindrical coordinates[2] (r,φ,z), (1) defines so-called r-analytic functions; see [30]. In the 3D *asymmetric* case, (1) relates the kth harmonics of ϕ and $\mathbf{\Lambda}$, $k \in \mathbb{Z}_0^+$, with respect to φ, and reduces to a series of systems of two linear first-order partial differential equations

$$\left(\frac{\partial}{\partial r} - \frac{k}{r}\right) U^{(k)} = \frac{\partial}{\partial z} V^{(k+1)}, \qquad \frac{\partial}{\partial z} U^{(k)} = -\left(\frac{\partial}{\partial r} + \frac{k+1}{r}\right) V^{(k+1)}, \tag{3}$$

[1] In contrast to generalized analytic functions, pseudoanalytic functions are defined axiomatically via generators [6].

[2] In this case, the z-axis is the axis of symmetry, and ϕ and $\mathbf{\Lambda}$ are independent of the angular coordinate φ.

which for each $k \in \mathbb{Z}_0^+$ defines a class of generalized analytic functions $G^{(k)}(r,z) = U^{(k)}(r,z) + i\,V^{(k+1)}(r,z)$ with $i = \sqrt{-1}$; see [30]. In particular, for $k = 0$, the system (3) defines the class of r-analytic functions.

It follows from (3) that $U^{(k)}$ and $V^{(k+1)}$ are k-harmonic and $(k+1)$-harmonic functions, respectively; i.e., they satisfy

$$\Delta_k U^{(k)} = 0 \quad \text{and} \quad \Delta_{k+1} V^{(k+1)} = 0, \tag{4}$$

where Δ_k denotes the so-called k-harmonic operator:

$$\Delta_k \equiv \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2} - \frac{k^2}{r^2}. \tag{5}$$

To emphasize this fact and in particular that (3) is associated with the kth harmonics of the related potentials and also to differentiate this class from other classes of generalized analytic (or pseudoanalytic) functions, we call the functions defined by the system (3) *k-harmonically analytic functions*.

Introducing the complex variable $\zeta = r + i\,z$, its conjugate $\overline{\zeta} = r - i\,z$ (whence $r = \frac{1}{2}(\zeta + \overline{\zeta})$ and $z = \frac{1}{2i}(\zeta - \overline{\zeta})$), and corresponding partial derivatives

$$\frac{\partial}{\partial \zeta} = \frac{1}{2}\left(\frac{\partial}{\partial r} - i\,\frac{\partial}{\partial z}\right), \qquad \frac{\partial}{\partial \overline{\zeta}} = \frac{1}{2}\left(\frac{\partial}{\partial r} + i\,\frac{\partial}{\partial z}\right),$$

we can represent the system (3) in the form

$$\frac{\partial G^{(k)}}{\partial \overline{\zeta}} = \frac{1}{4r}\left((2k+1)\,\overline{G^{(k)}} - G^{(k)}\right), \tag{6}$$

which has an advantage over (3) in certain formal manipulations. However, we should note that there is a difference in defining $G^{(k)}$ by (3) and (6): the derivative $\partial/\partial\overline{\zeta}$ may exist when $\partial/\partial r$ and $\partial/\partial z$ do not. Further, we will formally write $G^{(k)}(\zeta) = G^{(k)}(r,z)$ without assuming analyticity of $G^{(k)}$. It follows from (6) that a k-harmonically analytic function multiplied by a real-valued constant remains k-harmonically analytic, and the sum and difference of any two k-harmonically analytic functions are again k-harmonically analytic. The last property, however, does not hold for the product of two arbitrary k-harmonically analytic functions.

The class of 0-harmonically analytic functions was studied by Alexandrov and Soloviev [1] in application to axially symmetric problems of the linear theory of elasticity. In particular, they generalized Cauchy's integral formula and constructed an integral representation for 0-harmonically analytic functions via ordinary analytic functions for an arbitrary region. Also, for this class of functions, the Hilbert formulae were derived for the regions exterior to cyclidal bodies (spindle, lens, bispheres, and torus) and used for obtaining analytic expressions of the pressure in 3D axially symmetric Stokes flows about solid spindle, lens, bispheres, and torus [32, 31, 30].

As discussed in [30], the system (3) is a particular case of the Bers–Vekua system and also, in some sense, can be viewed as a particular case of Položii's system with $p = r^{2k+1}$, $U^{(k)} = r^k \widehat{U}$, and $V^{(k+1)} = r^{-(k+1)}\widehat{V}$. However, specializing results of the theories of generalized analytic (pseudoanalytic) and p-analytic functions to the class of k-harmonically analytic functions is not straightforward. For example, obtaining Cauchy's integral formula in the framework of the Bers–Vekua system for each particular subclass requires solving certain singular integral equations for finding so-called "main solutions," which is a nontrivial task. Also, the existing Cauchy integral formula for p-analytic functions with the characteristic $p = r^{2k+1}$ cannot be translated directly to the one for functions satisfying (3), since the dependence between $G^{(k)}$ and P is established only through their real and imaginary parts. On the other hand, the fact that $U^{(k)}$ and $V^{(k+1)}$ solve (4) makes constructing representations for k-harmonically analytic functions in canonical regions[3] significantly easier via (3) than doing so as a corollary of the aforementioned theories. In this paper, we construct an integral representation for k-harmonically analytic functions via ordinary analytic functions and derive series representations for k-harmonically analytic functions for the regions exterior to sphere and prolate and oblate spheroids. Using the approach of Alexandrov and Soloviev [1], we also obtain Cauchy's integral formula for k-harmonically analytic functions, which plays a central role in solving Stokes flow problems for noncanonical regions.

1.2. Stokes model in axially symmetric case. As the central result of the first part of the developed framework, we obtain a representation (solution form) for the velocity field and pressure for 3D *axially symmetric* Stokes flows in terms of two 0-harmonically analytic functions. The representation is similar to Goursat's formula, representing a solution to a 2D biharmonic equation via two ordinary analytic functions, and reduces axially symmetric Stokes flow problems to boundary-value problems for two 0-harmonically analytic functions. For the canonical regions, the boundary-value problems can be solved analytically based on representations of 0-harmonically analytic functions in corresponding curvilinear coordinates, and for regions in which Laplace's equation does not admit separation of variables, they can be reduced to an integral equation[4] based on Cauchy's integral formula for 0-harmonically analytic functions. This is the main advantage of the suggested solution form compared to the stream function approach [11] widely used for solving axially symmetric Stokes flow problems. We also express the resisting (drag) force, exerted on a solid body of revolution in the axially symmetric translation, in terms of a 0-harmonically analytic function entering the solution form. In the second part of the developed framework (see [29]), we construct a representation (solution form) for the velocity field and pressure for 3D *asymmetric* Stokes flows in terms of three k-harmonically analytic functions and demonstrate the developed framework in solving several asymmetric Stokes flow problems.

The rest of the paper is organized into four sections. Section 2 presents main results for k-harmonically analytic functions: integral representation via ordinary analytic functions, representations for the regions exterior to sphere and prolate and oblate spheroids, Cauchy's integral formula, and some auxiliary results. Section 3 constructs the solution form for axially symmetric Stokes flow problems in terms of two 0-harmonically analytic functions, proves that the solution form uniquely determines an external velocity field, expresses the resisting force in terms of a 0-harmonically analytic function entering the solution form, and derives the integral equation based on Cauchy's integral formula for 0-harmonically analytic functions. Section 4 demonstrates the solution form in obtaining closed-form analytical solutions to the 3D Stokes flow problem for the axially symmetric translation of solid sphere and solid prolate and oblate spheroids and solves the integral equation for the axially symmetric translation of solid bispheroids[5] and a solid torus of elliptical cross-section for various values of a

[3] By canonical regions we mean those in which Laplace's equation admits separation of variables.

[4] For discussion of integral equation approaches and standard numerical techniques in application to hydrodynamic problems, the reader may refer to [21, 22].

[5] Bispheroids are two separate spheroids of equal size with the same axis of revolution.

geometrical parameter. The appendix presents the proof of Cauchy's integral formula for k-harmonically analytic functions.

2. *k*-harmonically analytic functions.

This section presents several results for k-harmonically analytic functions which are central in application to 3D Stokes flow problems: an integral representation via ordinary analytic functions, Cauchy's integral formula, series representations for the regions exterior to sphere and prolate and oblate spheroids, and some auxiliary results.

Let $\zeta = r + i\,z$, and let $G^{(k)}(\zeta) = U^{(k)}(\zeta) + i\,V^{(k+1)}(\zeta)$ be a k-harmonically analytic function satisfying (3) in a region $\mathcal{D}$. In this work, we consider $\mathcal{D}$ to be symmetric with respect to the z-axis. In this case, the boundary ℓ of $\mathcal{D}$ consists of the right part ℓ_+ ($\operatorname{Re}\zeta \geq 0$) and the left part ℓ_- ($\operatorname{Re}\zeta \leq 0$), which is the symmetric reflection of ℓ_+ with respect to the z-axis. The parts ℓ_+ and ℓ_- can be either closed curves or open contours with the endpoints lying on the z-axis. Since $G^{(k)}(\zeta)$ has "physical" meaning only for $\operatorname{Re}\zeta \geq 0$, we can define the function $G^{(k)}(\zeta)$ for $\operatorname{Re}\zeta < 0$ by the symmetry condition

$$G^{(k)}\left(-\overline{\zeta}\right) = (-1)^k\overline{G^{(k)}(\zeta)}, \tag{7}$$

which is dictated by formulae representing the velocity field of Stokes flows in terms of k-harmonically analytic functions.

For the class of 0-harmonically analytic functions, Cauchy's integral formula and integral representation via ordinary analytic functions were obtained by Alexandrov and Soloviev [1], who also suggested (7) for $k = 0$. Also, some results for k-harmonically analytic functions can be adopted from the theory of p-analytic functions [20] through the relationship between $G^{(k)}(\zeta)$ and a p-analytic function $P(\zeta) = \widehat{U} + i\,\widehat{V}$ with the characteristic $p = r^{2k+1}$:

$$P(\zeta) = r^{-k}U^{(k)}(\zeta) + i\,r^{k+1}V^{(k+1)}(\zeta). \tag{8}$$

2.1. Representations of *k*-harmonically analytic functions.

The following proposition constructs the main integral representation for k-harmonically analytic functions, which generalizes the integral representation for 0-harmonically analytic functions [1, formula (28.2), p. 248] and is a modification of the existing integral representation for p-analytic functions with $p = r^{2k+1}$ (see [20, p. 177]).

Proposition 1 (main integral representation). *In a simply connected region $\mathcal{D}$, a k-harmonically analytic function $G^{(k)}(\zeta)$ can be represented in the form*

$$G^{(k)}(\zeta) = \frac{1}{r^k|r|}\int_{-\overline{\zeta}}^{\zeta} f(\tau)\,(\zeta-\tau)^{k-\frac{1}{2}}\left(\overline{\zeta}+\tau\right)^{k+\frac{1}{2}}d\tau, \tag{9}$$

where $f(\tau)$ is an analytic function of $\tau = r_1 + i\,z_1$ in the region $\mathcal{D}$ such that $f\left(-\overline{\tau}\right) = \overline{f(\tau)}$; and the points $-\overline{\zeta}$ and ζ are connected by an arbitrary simple curve C lying in $\mathcal{D}$ and symmetric with respect to the z-axis. If $\mathcal{D}$ is unbounded, then convergence of the integral (9) requires $f(\tau) \sim \mathcal{O}\left(|\tau|^{-2k-1-\epsilon}\right)$ with $\epsilon > 0$ at $|\tau| \to \infty$.

Proof. We use the relationship (8) and the existing representations for the real and imaginary parts of a p-analytic function (see [20, p. 178, formulae (48) and (49)]) with $p = r^{2k+1}$ to formally write

$$U^{(k)}(\zeta) = \frac{1}{r^{k-1}|r|}\int_{-\overline{\zeta}}^{\zeta} f(\tau)\left[(\zeta-\tau)(\overline{\zeta}+\tau)\right]^{k-\frac{1}{2}}d\tau, \tag{10}$$

$$i\,V^{(k+1)}(\zeta) = \frac{1}{r^k|r|}\int_{-\overline{\zeta}}^{\zeta} f(\tau)(\tau - i\,z)\left[(\zeta-\tau)(\overline{\zeta}+\tau)\right]^{k-\frac{1}{2}}d\tau, \tag{11}$$

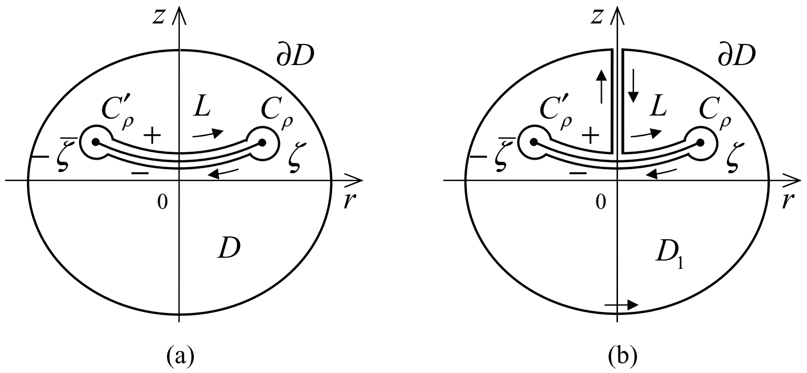
where a curve connecting ζ and $-\overline{\zeta}$ is simple and symmetric with respect to the z-axis. The representation (9) follows from (10) and (11). The absolute value of r in the multipliers at the integrals in (10) and (11) is introduced to satisfy the symmetry condition (7). We need to show that (9) is in fact a k-harmonically analytic function.

The function $\left[(\zeta-\tau)\left(\overline{\zeta}+\tau\right)\right]^{k-\frac{1}{2}}$ is analytic with respect to τ everywhere in the complex plane except for $\tau=\zeta$ and $\tau=-\overline{\zeta}$ and has two branches. With a branch cut connecting the points $-\overline{\zeta}$ and ζ, each branch is uniquely determined, and we choose the one which assumes nonnegative values at the upper bank of the branch cut (see Figure 1(a)). Consequently, at the upper and lower banks of the branch cut, the integral (9) has opposite signs, and we can equivalently represent it as the integral over the closed curve L consisting of parts of the upper and lower banks of the branch cut and of the circles C_ρ and C'_ρ with radius ρ and centers at ζ and $-\overline{\zeta}$, respectively (see Figure 1(a)):

$$G^{(k)}(\zeta)=\frac{1}{2r^k|r|}\int_L f(\tau)\,(\zeta-\tau)^{k-\frac{1}{2}}\left(\overline{\zeta}+\tau\right)^{k+\frac{1}{2}}d\tau. \tag{12}$$

Indeed, as $\rho\to 0$, the integrals of $g(\tau)=f(\tau)\,(\zeta-\tau)^{k-\frac{1}{2}}\left(\overline{\zeta}+\tau\right)^{k+\frac{1}{2}}$ over the circles C_ρ and C'_ρ for $k\geq 0$ vanish, and thus (12) reduces to the sum of the integrals along the upper and lower banks of the branch cut.

First, we show that (9) is independent of the curve C, connecting the points $-\overline{\zeta}$ and ζ within the region $\mathcal{D}$. Since C can be chosen to be the upper bank of the branch cut, we need to show that (9) is independent of the form of the branch cut. To this end, we use the representation (12) and make a crosscut in the region D connecting the closed curve L with the boundary ∂D and forming a simply connected region D_1; see Figure 1(b). In D_1, the function $g(\tau)$ is analytic, and, consequently, its integral over the boundary of D_1 vanishes. This means that the integral of $g(\tau)$ over L with an arbitrary branch cut in D equals the integral of $g(\tau)$ over ∂D with the opposite orientation.



(a) (b)

FIG. 1. (a) *Closed curve L consisting of parts of the upper and lower banks of the branch cut, connecting the points $-\overline{\zeta}$ and ζ, and of the circles C_ρ and C'_ρ with radius ρ and centers at ζ and $-\overline{\zeta}$, respectively;* (b) *The region D_1 is formed by connecting L with ∂D via a crosscut.*

Next, we show that (9) has continuous partial derivatives $\partial/\partial r$ and $\partial/\partial z$ in $\mathcal{D}$, in particular at $r=0$. We represent (9) in the form of (12). Since the curve L in (12) contains neither ζ nor $-\overline{\zeta}$, the integral (12) is continuously differentiable with respect to r and z. For the case of $r \to 0$, let ζ be close to the z-axis and let the branch cut connecting $-\overline{\zeta}$ and ζ be a segment, i.e., $\tau = r_1 + i\,z$, $r_1 \in [-r, r]$, and $d\tau = dr_1$. Using the condition $f\left(-\overline{\tau}\right) = \overline{f(\tau)}$ and the change of variable $\tau(x) = r\sin x + i\,z$, $x \in [0, \pi/2]$, we obtain

$$G^{(k)}(\zeta) = 2\,r^k \int_0^{\frac{\pi}{2}} \left(\mathrm{Re}\left[f\left(r\sin x + i\,z\right)\right] + i\,\sin x\;\mathrm{Im}\left[f\left(r\sin x + i\,z\right)\right]\right)(\cos x)^{2k}\,dx, \tag{13}$$

for which the derivatives $\partial/\partial r$ and $\partial/\partial z$ at $r = 0$ obviously exist and are continuous.

Finally, we show that (12) satisfies the Carleman system (6). Since the partial derivatives $\partial/\partial r$ and $\partial/\partial z$ of (9) are continuous in $\mathcal{D}$, the derivative $\partial/\partial\overline{\zeta}$ of (9) is also continuous in $\mathcal{D}$ and for $r > 0$ it takes the form

$$\frac{\partial}{\partial\overline{\zeta}} G^{(k)}(\zeta) = \frac{1}{4r}\left((-1)^k\,(2k+1)\,G^{(k)}\left(-\overline{\zeta}\right) - G^{(k)}(\zeta)\right). \tag{14}$$

Since $f\left(-\overline{\tau}\right) = \overline{f(\tau)}$, we have $G^{(k)}\left(-\overline{\zeta}\right) = (-1)^k\overline{G^{(k)}(\zeta)}$, which is the symmetry condition (7), and thus with (7), (14) is equivalent to (6). □

For $k = 0$, the formula (9) coincides with the one presented by Alexandrov and Soloviev [1] (with a necessary modification due to the difference in notation[6]).

We illustrate (9) in obtaining a series representation for a k-harmonically analytic function for the region exterior to a sphere. Let (R, ϑ, φ) be the spherical coordinates related to the cylindrical coordinates (r, φ, z) by $r = R\sin\vartheta$ and $z = R\cos\vartheta$ with $R \in \mathbb{R}_0^+$ and $\vartheta \in [0, \pi]$.

Example 2 (region exterior to a sphere). For the region exterior to a sphere, a k-harmonically analytic function $G^{(k)}$ is represented by

$$\begin{aligned} G^{(k)}(R,\vartheta) &= U^{(k)}(R,\vartheta) + i\,V^{(k+1)}(R,\vartheta) \\ &= \sum_{n=1}^{\infty} A_n R^{-n-k-1}\left\{n\,\mathrm{P}_{n+k}^{(k)}(\cos\vartheta) - i\,\mathrm{P}_{n+k}^{(k+1)}(\cos\vartheta)\right\}, \end{aligned} \tag{15}$$

where A_n, $n \geq 1$, is a real-valued constant, and $\mathrm{P}_m^{(k)}(\cos\vartheta)$ is the associated Legendre polynomial of the first kind of order m and rank k (for $k = 0$, the superscript is omitted). The behavior of A_n at $n \to \infty$ follows from the requirements for (15) to converge.

Detail. The region $\mathcal{D}$ exterior to a sphere with the branch cut along the ray $\vartheta = \pi$ is simply connected. For $\mathcal{D}$, an analytic function $f(\zeta)$, vanishing as $|\zeta|^{-2k-1-\epsilon}$, $\epsilon > 0$, at $|\zeta| \to \infty$ and satisfying the condition $f\left(-\overline{\zeta}\right) = \overline{f(\zeta)}$, can be represented by the series

$$f(\zeta) = \sum_{n=2(k+1)}^{\infty} a_n(-i\zeta)^{-n},$$

where a_n, $n \geq 1$, is a real-valued constant.

According to (10) and (11), for $r \geq 0$, the real and imaginary parts of a k-harmonically analytic function are represented by

$$U^{(k)}(\zeta) = \sum_{n=2(k+1)}^{\infty} a_n I_n(\zeta), \qquad i\,V^{(k+1)}(\zeta) = \sum_{n=2(k+1)}^{\infty} a_n J_n(\zeta), \tag{16}$$

[6] In [1], the complex variable ζ is introduced as $\zeta = z + i\,r$.

where

$$I_n(\zeta) = r^{-k} \int_{-\overline{\zeta}}^{\zeta} (-i\tau)^{-n} \left[(\zeta - \tau)(\overline{\zeta} + \tau)\right]^{k-\frac{1}{2}} d\tau,$$

$$J_n(\zeta) = r^{-k-1} \int_{-\overline{\zeta}}^{\zeta} (-i\tau)^{-n} (\tau - i\, z) \left[(\zeta - \tau)(\overline{\zeta} + \tau)\right]^{k-\frac{1}{2}} d\tau.$$

To calculate the integrals $I_n(\zeta)$ and $J_n(\zeta)$, let a curve connecting ζ and $-\overline{\zeta}$ be the arc of the circle with the radius $|\zeta|$ and center at $\zeta = 0$. In this case, in the spherical coordinates, we have $\zeta = i\, R\, \mathrm{e}^{-i\vartheta}$ and $\tau = i\, R\, \mathrm{e}^{-i\varphi}$, and thus, the arc is parameterized by $\varphi \in [-\vartheta, \vartheta]$ and does not intersect the ray $\vartheta = \pi$. Consequently, $(\zeta - \tau)(\overline{\zeta} + \tau) = 2R^2 \mathrm{e}^{-i\varphi}(\cos\varphi - \cos\vartheta)$, and the integrals I_n and J_n reduce to

$$\begin{aligned} I_n(R, \vartheta) &= 2^{k+\frac{1}{2}} R^{k-n} (\sin\vartheta)^{-k} \int_0^{\vartheta} \cos\left[\left(k + \tfrac{1}{2} - n\right) \varphi\right] (\cos\varphi - \cos\vartheta)^{k-\frac{1}{2}} d\varphi \\ &= (-2)^k \sqrt{\pi}\, \frac{\Gamma\left(k + \frac{1}{2}\right) \Gamma(n - 2k)}{\Gamma(n)} R^{k-n}\, \mathrm{P}_{n-k-1}^{(k)}(\cos\vartheta), \end{aligned} \tag{17}$$

$$\begin{aligned} J_n(R, \vartheta) = i\, 2^{k+\frac{1}{2}} R^{k-n} (\sin\vartheta)^{-k-1} \Bigg\{ &\int_0^{\vartheta} \cos\left[\left(k + \tfrac{1}{2} - n\right) \varphi\right] (\cos\varphi - \cos\vartheta)^{k+\frac{1}{2}} d\varphi \\ &- \int_0^{\vartheta} \sin\varphi \sin\left[\left(k + \tfrac{1}{2} - n\right) \varphi\right] (\cos\varphi - \cos\vartheta)^{k-\frac{1}{2}} d\varphi \Bigg\} \\ = -i\, (-2)^k \sqrt{\pi}\, &\frac{\Gamma\left(k + \frac{1}{2}\right) \Gamma(n - 2k - 1)}{\Gamma(n)} R^{k-n}\, \mathrm{P}_{n-k-1}^{(k+1)}(\cos\vartheta), \end{aligned} \tag{18}$$

where $\Gamma(\cdot)$ is the Gamma function, and we used the representation for the associated Legendre polynomial

$$\int_0^{\vartheta} \cos\left[\left(k + \tfrac{1}{2} - n\right) \varphi\right] (\cos\varphi - \cos\vartheta)^{k-\frac{1}{2}} d\varphi = \sqrt{\frac{\pi}{2}}\, (\sin\vartheta)^k\, \Gamma\left(k + \tfrac{1}{2}\right) \mathrm{P}_{n-k-1}^{(-k)}(\cos\vartheta);$$

see [3]. Denoting $A_n = (-2)^k \sqrt{\pi}\, \frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma(n)}{\Gamma(n+2k+1)}\, a_{n+2k+1}$, and substituting (17) and (18) into (16), we obtain (15).

Another approach for deriving representations for the k-harmonically analytic function in a canonical region, i.e., in which Laplace's equation admits separation of variables, e.g., prolate and oblate spheroids, two spheres, torus, lens, spindle, etc., is based on Hilbert formulae [30, 31, 32] and usually requires less effort compared to that of making use of the general representation (9). This approach represents solutions to (4) in the form of integrals or series in curvilinear coordinates, associated with the geometry of the region, and finds a relationship between those integrals or series based on the system (3). It is shown in [30] that for the functions satisfying (4) and vanishing at infinity (in an outer region), it is sufficient to use only one equation of (3).

The next example illustrates this approach for the region exterior to a prolate spheroid. To this end, we introduce prolate spheroidal coordinates (ξ, η, φ) related to the cylindrical coordinates by

$$r = c \sinh\xi\, \sin\eta, \quad z = c \cosh\xi\, \cos\eta, \qquad \xi \in [0, \infty), \quad \eta \in [0, \pi], \tag{19}$$

where the angular coordinate $\varphi \in [0, 2\pi)$ coincides with the one in (r, φ, z) and c is a metric parameter. A prolate spheroid is determined by fixing the coordinate ξ, i.e., $\xi = \xi_0$.

Example 3 (region exterior to a prolate spheroid). For the region exterior to the prolate spheroid ($\xi \geq \xi_0$), a k-harmonically analytic function $G^{(k)}$ is represented by

$$\begin{aligned} G^{(k)}(\xi,\eta) &= U^{(k)}(\xi,\eta) + i\,V^{(k+1)}(\xi,\eta) \\ &= \sum_{n=0}^{\infty} A_n \Big\{ (n-k)(n+k+1)\,\mathrm{Q}_n^{(k)}(\cosh\xi)\,\mathrm{P}_n^{(k)}(\cos\eta) \\ &\qquad + i\,\mathrm{Q}_n^{(k+1)}(\cosh\xi)\,\mathrm{P}_n^{(k+1)}(\cos\eta) \Big\}, \end{aligned} \tag{20}$$

where A_n, $n \geq 1$, are real-valued constants and $\mathrm{Q}_n^{(k)}(\cosh\xi)$ is the associated Legendre function of the second kind.

Detail. It is known that the functions $U^{(k)}(\xi,\eta)$ and $V^{(k+1)}(\xi,\eta)$ that satisfy (4) in the region exterior to the prolate spheroid can be represented in the form of series (see [12, 16])

$$\begin{aligned} U^{(k)}(\xi,\eta) &= \sum_{n=0}^{\infty} a_n\,\mathrm{Q}_n^{(k)}(\cosh\xi)\,\mathrm{P}_n^{(k)}(\cos\eta), \\ V^{(k+1)}(\xi,\eta) &= \sum_{n=0}^{\infty} b_n\,\mathrm{Q}_n^{(k+1)}(\cosh\xi)\,\mathrm{P}_n^{(k+1)}(\cos\eta), \end{aligned} \tag{21}$$

where a_n and b_n are real-valued coefficients. In the prolate spheroidal coordinates (19), the system (3) takes the form

$$\begin{aligned} \left(\frac{\partial}{\partial\xi} - k\coth\xi\right) U^{(k)} &= -\left(\frac{\partial}{\partial\eta} + (k+1)\cot\eta\right) V^{(k+1)}, \\ \left(\frac{\partial}{\partial\eta} - k\cot\eta\right) U^{(k)} &= \left(\frac{\partial}{\partial\xi} + (k+1)\coth\xi\right) V^{(k+1)}. \end{aligned}$$

Substituting the series (21) into any equation of the system above, we obtain the relationship for a_n and b_n:

$$a_n = (n-k)(n+k+1)\,b_n.$$

Consequently, with $A_n = b_n$, the representation (20) follows.

The following example presents a k-harmonically analytic function in the form of series for the region exterior to an oblate spheroid. In the oblate spheroidal coordinates (ξ, η, φ) related to the cylindrical coordinates by

$$r = c\,\cosh\xi\,\sin\eta, \qquad z = c\,\sinh\xi\,\cos\eta, \qquad \xi \in [0,\infty), \quad \eta \in [0,\pi], \tag{22}$$

with the same angular coordinate φ (c is a metric parameter), the oblate spheroid is determined by fixing the coordinate ξ, i.e., $\xi = \xi_0$.

Example 4 (region exterior to an oblate spheroid). For the region exterior to the *oblate* spheroid ($\xi \geq \xi_0$), a k-harmonically analytic function $G^{(k)}$ is represented by

$$\begin{aligned} G^{(k)}(\xi,\eta) &= U^{(k)}(\xi,\eta) + i\,V^{(k+1)}(\xi,\eta) \\ &= \sum_{n=0}^{\infty} A_n\, i^{n+1} \Big\{ (n-k)(n+k+1)\,\mathrm{Q}_n^{(k)}(i\sinh\xi)\,\mathrm{P}_n^{(k)}(\cos\eta) \\ &\qquad + i\,\mathrm{Q}_n^{(k+1)}(i\sinh\xi)\,\mathrm{P}_n^{(k+1)}(\cos\eta) \Big\}, \end{aligned} \tag{23}$$

where A_n, $n \geq 1$, are real-valued constants and $i^{n+1}\,\mathrm{Q}_n^{(k)}(i\sinh\xi)$ and $i^{n+1}\,\mathrm{Q}_n^{(k+1)}(i\sinh\xi)$ are real-valued functions.

Detail. In the oblate spheroidal coordinates (22), the system (3) takes the form

$$\left(\frac{\partial}{\partial\xi} - k\tanh\xi\right)U^{(k)} = -\left(\frac{\partial}{\partial\eta} + (k+1)\cot\eta\right)V^{(k+1)},$$
$$\left(\frac{\partial}{\partial\eta} - k\cot\eta\right)U^{(k)} = \left(\frac{\partial}{\partial\xi} + (k+1)\tanh\xi\right)V^{(k+1)},$$

and the representation (23) is obtained similarly to (20).

Representations for k-harmonically analytic functions for the regions exterior to bispheres, torus, lens, and spindle are discussed in [32, 31, 30].

2.2. Cauchy's integral formula for *k*-harmonically analytic functions. In the previous section, we have transformed the integral representation for p-analytic functions with $p = r^{2k+1}$ into the integral representation (9) for k-harmonically analytic functions. However, doing the same for the existing Cauchy integral formula for p-analytic functions is not a straightforward task, since the relationship between p-analytic and k-harmonically analytic functions is established only via their real and imaginary parts. To derive Cauchy's integral formula for k-harmonically analytic functions, we use the approach of Alexandrov and Soloviev [1], who obtained Cauchy's integral formula for the class of 0-harmonically analytic functions.

In the rz-plane, let $\mathcal{D}^+$ be a bounded (inner) region symmetric with respect to the z-axis, and let $\mathcal{D}^-$ be the outer region with respect to $\mathcal{D}^+$ (i.e., the complement of int$\mathcal{D}^+$). $\mathcal{D}_0^+$ and $\mathcal{D}_0^-$ will denote $\mathcal{D}^+$ with $r \geq 0$ and $\mathcal{D}^-$ with $r \geq 0$, respectively, i.e., the right parts of the corresponding regions. Let ℓ be the common boundary of $\mathcal{D}^+$ and $\mathcal{D}^-$, and let ℓ_+ denote the right part of ℓ (ℓ for $r \geq 0$), which is either a closed curve or an open curve with the endpoints lying on the z-axis. Thus, $\ell = \ell_+ \bigcup \ell_-$, where ℓ_- is the reflection of ℓ_+ with respect to the z-axis. The boundary ℓ is positively oriented or traversed in the counterclockwise direction if $\mathcal{D}^+$ remains on the left side when one travels along ℓ in this direction.[7]

THEOREM 2 (Cauchy's integral formula for k-harmonically analytic functions). *Let $\mathcal{D}^+$ be a simply connected region with the smooth, positively oriented boundary $\ell = \ell_+ \bigcup \ell_-$ (symmetric with respect to the z-axis), and let $G^{(k)}(\zeta)$ be a k-harmonically analytic function in $\mathcal{D}_0^+$ satisfying the symmetry condition* (7) *and the Hölder condition*[8] *on ℓ. Cauchy's integral formula for the function $G^{(k)}(\zeta)$ is given by*

(24)
$$G^{(k)}(\zeta) = \frac{1}{2\pi i}\oint_\ell G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\,d\tau$$
$$\equiv \frac{1}{2\pi i}\int_{\ell_+}\left(G^{(k)}(\tau)\,\Omega_+^{(k)}(\zeta,\tau)\,\frac{d\tau}{\tau-\zeta} - \overline{G^{(k)}(\tau)}\,\Omega_-^{(k)}(\zeta,\tau)\,\frac{d\overline{\tau}}{\overline{\tau}+\zeta}\right), \quad \zeta \in \operatorname{int}\mathcal{D}^+,$$

[7] The orientation of a closed curve is always determined with respect to the corresponding inner region.

[8] This condition means that for some parametrization $\zeta(t)$ of the curve ℓ, the boundary value $G^{(k)}(\zeta(t))$ satisfies $|G^{(k)}(\zeta(t_2)) - G^{(k)}(\zeta(t_1))| \leq C\,|t_2 - t_1|^\beta$ for all t_1 and t_2, some $\beta \in (0,1]$, and nonnegative constant C.

where $\Omega_+^{(k)}(\zeta,\tau)$ *and* $\Omega_-^{(k)}(\zeta,\tau)$ *are real-valued functions determined by*

(25a)

$$\begin{aligned}\Omega_+^{(k)}(\zeta,\tau) = \frac{\left[\Gamma\left(k+\frac{3}{2}\right)\right]^2}{\Gamma(2(k+1))}\left|\frac{\tau+\overline{\tau}}{\zeta+\overline{\tau}}\right|[\lambda(\zeta,\tau)]^{2k}\\ \times\left(1-\lambda^2(\zeta,\tau)\right)\mathbb{F}\left(k+\tfrac{3}{2},k+\tfrac{3}{2},2(k+1),\lambda^2(\zeta,\tau)\right),\end{aligned}$$

(25b)

$$\Omega_-^{(k)}(\zeta,\tau) = \frac{\left[\Gamma\left(k+\frac{3}{2}\right)\right]^2}{\Gamma(2(k+1))}\left|\frac{\tau+\overline{\tau}}{\zeta+\overline{\tau}}\right|[\lambda(\zeta,\tau)]^{2k}\,\mathbb{F}\left(k+\tfrac{1}{2},k+\tfrac{3}{2},2(k+1),\lambda^2(\zeta,\tau)\right),$$

in which $\mathbb{F}(a,b,c,\kappa)$ *is the hypergeometric function*

$$\mathbb{F}(a,b,c,\kappa) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-\kappa\, t)^a}dt \tag{26}$$

and

$$\lambda^2(\zeta,\tau) = \frac{\left|\left(\zeta+\overline{\zeta}\right)\left(\tau+\overline{\tau}\right)\right|}{\left|\zeta+\overline{\tau}\right|^2}. \tag{27}$$

If in the above conditions $G^{(k)}(\zeta)$ *is instead k-harmonically analytic in* $\mathcal{D}_0^-$ *and vanishing at infinity, then* (24) *holds for* $\zeta\in\operatorname{int}\mathcal{D}^-$ *with negatively oriented* ℓ *(with respect to* $\mathcal{D}^+$*).*

Proof. The proof is presented in the appendix. □

We should note that $\frac{1}{2\pi i}\oint_\ell G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\,d\tau$ in Cauchy's integral formula (24) should be understood as an operator $\mathcal{A}\left(G^{(k)};\ell\right)$ rather than an ordinary integral. For example, it follows from the second line in (24) that $\mathcal{A}\left(c\,G^{(k)};\ell\right)\neq c\,\mathcal{A}\left(G^{(k)};\ell\right)$ for an arbitrary complex-valued constant c, whereas this fact does not come from the properties of integrals.

Remark 1. Cauchy's integral formula (24) holds for multiply connected regions. The proof is similar to that for Cauchy's integral formula for ordinary analytic functions in the case of a multiply connected region.

Remark 2. In Cauchy's integral formula (24), while $\lim_{\tau\to\zeta}\Omega_+^{(k)}(\zeta,\tau)=1$, the function $\Omega_-^{(k)}(\zeta,\tau)$ has a logarithmic singularity at $\tau=\zeta$:

(28)

$$\Omega_-^{(k)}(\zeta,\tau) = -(2k+1)\left(\ln\frac{|\tau-\zeta|}{|\zeta+\overline{\tau}|}+\frac{1}{2k+1}+\gamma+\psi\left(k+\tfrac{1}{2}\right)\right)+\mathcal{O}(|\tau-\zeta|)\ \text{ as } \tau\to\zeta,$$

where γ is the Euler constant and $\psi(\cdot)$ is the digamma function. This result follows from the asymptotic representation of $\mathbb{F}\left(k+1/2,k+3/2,2(k+1),\lambda^2(\zeta,\tau)\right)$ as $\lambda\to 1-$ (see the appendix). Note that there is no singularity at $\tau=\zeta$ if $\operatorname{Re}\zeta=0$.

Corollary 3 (Cauchy's integral formula for 0-harmonically analytic functions). *In the case of* 0*-harmonically analytic functions,* $\Omega_+^{(0)}(\zeta,\tau)$ *and* $\Omega_-^{(0)}(\zeta,\tau)$ *simplify (see also* [1]*):*

$$\Omega_+^{(0)}(\zeta,\tau) = \left|\frac{\tau+\overline{\zeta}}{\zeta+\overline{\zeta}}\right|\left\{\left(\lambda^2(\zeta,\tau)-1\right)\mathbb{K}(\lambda(\zeta,\tau))+\mathbb{E}(\lambda(\zeta,\tau))\right\},$$

$$\Omega_-^{(0)}(\zeta,\tau) = \left|\frac{\tau+\overline{\zeta}}{\zeta+\overline{\zeta}}\right|\left\{\mathbb{K}(\lambda(\zeta,\tau))-\mathbb{E}(\lambda(\zeta,\tau))\right\},$$

where $\mathbb{K}(\lambda(\zeta,\tau))$ and $\mathbb{E}(\lambda(\zeta,\tau))$ are complete elliptic integrals of the first and second kinds, respectively[9]*:*

$$\mathbb{K}(\lambda(\zeta,\tau)) = \int_0^{\frac{\pi}{2}} \frac{dt}{\sqrt{1-\lambda^2(\zeta,\tau)\sin^2 t}}, \qquad \mathbb{E}(\lambda(\zeta,\tau)) = \int_0^{\frac{\pi}{2}} \sqrt{1-\lambda^2(\zeta,\tau)\sin^2 t}\, dt.$$

The function $\mathbb{K}(\lambda)$ has a logarithmic singularity as $\lambda \to 1-$: $\mathbb{K}(\lambda) = -\frac{1}{2}\ln\left(1-\lambda^2\right) + \mathcal{O}(1)$.

Remark 3 (behavior at infinity). Let $\zeta = i\,R\,\mathrm{e}^{-i\vartheta}$, $\vartheta \in [0,\pi]$, in the spherical coordinates (R,ϑ,φ). At $|\zeta| \to \infty$, the asymptotic form for a k-harmonically analytic function, represented by Cauchy's integral formula (24), is determined by

$$G^{(k)}(\zeta) = \frac{1}{2\pi}\,(2k+1)\,\mathbb{B}\left(k+\tfrac{3}{2}, k+\tfrac{1}{2}\right) \mathrm{Re}\left[\int_{\ell_+} G^{(k)}(\tau)\,|\tau+\overline{\tau}|^{k+1}\,d\tau\right] \times \frac{(2\sin\vartheta)^k\,\mathrm{e}^{i\vartheta}}{|\zeta|^{k+2}} + \mathcal{O}\left(|\zeta|^{-(k+3)}\right), \qquad |\zeta| \to \infty,$$

where $\mathbb{B}\left(k+\frac{3}{2}, k+\frac{1}{2}\right)$ is the beta function.

Detail. Since $|\zeta| \to \infty$, we can represent

$$\frac{1}{\tau-\zeta} = -\frac{\zeta^{-1}}{1-\frac{\tau}{\zeta}} = -\sum_{n=0}^{\infty}\frac{\tau^n}{\zeta^{n+1}}, \qquad \frac{1}{\overline{\tau}+\zeta} = \frac{\zeta^{-1}}{1+\frac{\overline{\tau}}{\zeta}} = \sum_{n=0}^{\infty}(-1)^n\frac{\overline{\tau}^n}{\zeta^{n+1}}. \tag{29}$$

When $|\zeta| \to \infty$, we have $\lambda^2(\zeta,\tau) \sim 2\,|\tau+\overline{\tau}|\sin\vartheta/|\zeta|$ and

$$\begin{aligned} \Omega_+^{(k)}(\zeta,\tau) &= \left(k+\frac{1}{2}\right)\mathbb{B}\left(k+\tfrac{3}{2}, k+\tfrac{1}{2}\right)\frac{(2\sin\vartheta)^k\,|\tau+\overline{\tau}|^{k+1}}{|\zeta|^{k+1}} + \mathcal{O}\left(|\zeta|^{-(k+2)}\right), \\ \Omega_-^{(k)}(\zeta,\tau) &= \left(k+\frac{1}{2}\right)\mathbb{B}\left(k+\tfrac{3}{2}, k+\tfrac{1}{2}\right)\frac{(2\sin\vartheta)^k\,|\tau+\overline{\tau}|^{k+1}}{|\zeta|^{k+1}} + \mathcal{O}\left(|\zeta|^{-(k+2)}\right). \end{aligned} \tag{30}$$

Substituting (29) and (30) into (24), we obtain the statement of the remark.

2.3. Auxiliary results for k-harmonically analytic functions.

This section presents several auxiliary results for k-harmonically analytic functions.

Proposition 4. *Let $G^{(k)}(\zeta)$, $G_1^{(k)}(\zeta)$, and $G_2^{(k)}(\zeta)$ be k-harmonically analytic functions; then*

$$\text{(i)} \qquad \frac{\partial}{\partial\overline{\zeta}}\left(rG^{(k)}\right) = \frac{1}{4}\left((2k+1)\,\overline{G^{(k)}} + G^{(k)}\right),$$

$$\text{(ii)} \qquad \frac{\partial}{\partial\overline{\zeta}}\left(rG_1^{(k)}G_2^{(k)}\right) = \frac{1}{4}\,(2k+1)\left(\overline{G_1^{(k)}}G_2^{(k)} + G_1^{(k)}\overline{G_2^{(k)}}\right);$$

$$\textit{in particular,} \quad \frac{\partial}{\partial\overline{\zeta}}\left(r\left[G^{(k)}\right]^2\right) = \frac{1}{2}\,(2k+1)\,\overline{G^{(k)}}G^{(k)}.$$

Proof. Formulae (i) and (ii) follow from (6). □

[9] In Wolfram Research's Mathematica, $\mathbb{K}(\lambda)$ and $\mathbb{E}(\lambda)$ are defined by $\mathbb{K}(\lambda) = \int_0^{\frac{\pi}{2}} \frac{dt}{\sqrt{1-\lambda\sin^2 t}}$ and $\mathbb{E}(\lambda) = \int_0^{\frac{\pi}{2}} \sqrt{1-\lambda\sin^2 t}\,dt$.

PROPOSITION 5. *Let L be a piecewise smooth, positively oriented curve bounding a simply connected region $\mathcal{D}$, and let $\Phi(\zeta) = U(\zeta)+i\,V(\zeta)$ be a complex-valued function with continuous first-order partial derivatives in $\mathcal{D}$. Then*

$$\oint_L \Phi\,d\zeta = 2i \iint_{\mathcal{D}} \frac{\partial \Phi}{\partial \overline{\zeta}}\,drdz. \tag{31}$$

Proof. Using Green's theorem, we obtain

$$\begin{aligned}\oint_L \Phi\,d\zeta &= \oint_L (U dr - V dz) + i \oint_L (V dr + U dz)\\ &= -\iint_{\mathcal{D}} \left(\frac{\partial V}{\partial r} + \frac{\partial U}{\partial z}\right) drdz + i \iint_{\mathcal{D}} \left(\frac{\partial U}{\partial r} - \frac{\partial V}{\partial z}\right) drdz\\ &= -2 \iint_{\mathcal{D}} \operatorname{Im}\left[\frac{\partial \Phi}{\partial \overline{\zeta}}\right] drdz + 2i \iint_{\mathcal{D}} \operatorname{Re}\left[\frac{\partial \Phi}{\partial \overline{\zeta}}\right] drdz = 2i \iint_{\mathcal{D}} \frac{\partial \Phi}{\partial \overline{\zeta}}\,drdz,\end{aligned}$$

and the proposition is proved. □

PROPOSITION 6. *Let $G^{(0)}(\zeta)$ be a 0-harmonically analytic function, and let ℓ_+ be a simple, positively oriented, piecewise smooth curve in the right half of the rz-plane (ℓ_+ either is closed or has the endpoints lying on the z-axis). Then*

$$\operatorname{Re}\left[\int_{\ell_+} r\,G^{(0)}(\zeta)\,d\zeta\right] = -2 \lim_{z\to\infty} \left(z^2 \operatorname{Re} G^{(0)}(r,z)\Big|_{r=0}\right). \tag{32}$$

Proof. Let $\mathcal{D}_R$ be a region in the right half of the rz-plane with a piecewise smooth, *negatively* oriented boundary L, which consists of ℓ_+, segments of the z-axis, and semicircle L_R with large radius R (see Figure 2). Using Proposition 5 and formula (i) in Proposition 4, we have

$$\oint_L r\,G^{(0)}(\zeta)\,d\zeta = -i \iint_{\mathcal{D}_R} \operatorname{Re}\left[G^{(0)}(r,z)\right] drdz.$$

Taking the real part of this equality and using the fact that the function $r\,G^{(0)}$ vanishes at the z-axis, we obtain

$$\operatorname{Re}\left[\int_{\ell_+} r\,G^{(0)}(\zeta)\,d\zeta + \int_{L_R} r\,G^{(0)}(\zeta)\,d\zeta\right] = 0. \tag{33}$$

The integral of $r\,G^{(0)}$ over L_R can be determined as follows. Let (R,ϑ,φ) be the spherical coordinates related to the cylindrical coordinates (r,φ,z) in the ordinary way, and let the function $G^{(0)}$ be represented in the spherical coordinates by (15) for $k=0$, i.e.,

$$G^{(0)}(R,\vartheta) = \sum_{n=1}^{\infty} A_n R^{-n-1} \left\{n\,\mathrm{P}_n(\cos\vartheta) - i\,\mathrm{P}_n^{(1)}(\cos\vartheta)\right\} = A_1\,R^{-2} \mathrm{e}^{i\vartheta} + \mathcal{O}\left(R^{-3}\right). \tag{34}$$

For L_R with negative (clockwise) orientation, we have $\zeta = i\,R\,\mathrm{e}^{-i\vartheta}$, where $\vartheta \in [0,\pi]$, and

$$\int_{L_R} r\,G^{(0)}(\zeta)\,d\zeta = 2\,A_1 + \mathcal{O}(R^{-1}), \tag{35}$$

where A_1 can be expressed from (34) as $A_1 = \lim_{z\to\infty}\left(z^2 \operatorname{Re} G^{(0)}(r,z)\big|_{r=0}\right)$. Passing R to infinity in (35) and using (33), we obtain (32). □

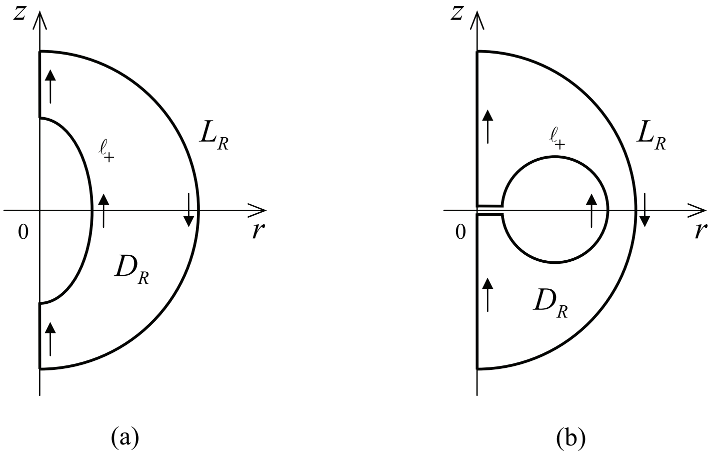


FIG. 2. *The region $\mathcal{D}_R$ with the piecewise smooth negatively oriented boundary L in two cases:* (a) *ℓ_+ is an open curve with the endpoints lying on the z-axis; and* (b) *ℓ_+ is a closed curve.*

3. Axially symmetric Stokes flows. In the axially symmetric case, in the cylindrical coordinates (r, φ, z) with the basis $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$ and the z-axis of symmetry, the velocity components u_r and u_z and the pressure $\wp$ for Stokes flows governed by (2) are independent of the angular coordinate φ, i.e.,

$$\mathbf{u} = u_r(r, z)\,\mathbf{e}_r + u_z(r, z)\,\mathbf{k}, \qquad u_\varphi \equiv 0, \qquad \wp = \wp(r, z).$$

To simplify notation, we will write a function of the variables of r and z as the function of ζ without assuming analyticity; e.g., $u_r(r, z)$ and $u_z(r, z)$ will be represented as $u_r(\zeta)$ and $u_z(\zeta)$, respectively.

A central result of this work is the representation of an external, axially symmetric velocity field and pressure that vanish at infinity in terms of two 0-harmonically analytic functions. This representation is similar to Goursat's formula with two ordinary analytic functions for a 2D biharmonic equation.

PROPOSITION 7 (representation of axially symmetric velocity field). *Let $G_1^{(0)}(\zeta) = U_1^{(0)}(\zeta) + i\,V_1^{(1)}(\zeta)$ and $G_2^{(0)}(\zeta) = U_2^{(0)}(\zeta) + i\,V_2^{(1)}(\zeta)$ be 0-harmonically analytic functions vanishing at infinity. Then in the axially symmetric case of the Stokes equations, the components u_r and u_z of an external velocity field vanishing at infinity can be represented in the form*

$$u_z(\zeta) + i\,u_r(\zeta) = \left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta), \tag{36}$$

and the pressure and vorticity also vanishing at infinity are determined by

$$\wp(\zeta) = \mu\,\mathrm{Re}\,G_1^{(0)}(\zeta), \qquad \mathrm{curl}\,\mathbf{u} = \mathrm{Im}\,G_1^{(0)}(\zeta)\;\mathbf{e}_\varphi. \tag{37}$$

Proof. In the axially symmetric case, the Stokes equations (2) reduce to

$$\mu\,\Delta_1 u_r = \frac{\partial \wp}{\partial r}, \qquad \mu\,\Delta_0 u_z = \frac{\partial \wp}{\partial z}, \tag{38}$$

and the continuity equation $\mathrm{div}\,\mathbf{u} = 0$ takes the form

$$\left(\frac{\partial}{\partial r} + \frac{1}{r}\right) u_r + \frac{\partial}{\partial z} u_z = 0, \tag{39}$$

where Δ_k is defined by (5). Our aim is to represent the velocity components u_r and u_z by linear combinations of 0-harmonically analytic functions (avoiding derivatives).

It follows from the Stokes equations (2) that the vorticity $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$ satisfies $\boldsymbol{\Delta\omega} = 0$. In the axially symmetric case, $\boldsymbol{\omega}$ takes the form $\boldsymbol{\omega} = \left(\frac{\partial u_r}{\partial z} - \frac{\partial u_z}{\partial r}\right) \mathbf{e}_\varphi$, and the equation $\boldsymbol{\Delta\omega} = 0$ reduces to $\Delta_1 \left(\frac{\partial u_r}{\partial z} - \frac{\partial u_z}{\partial r}\right) = 0$. Consequently, $\boldsymbol{\omega}$ can be represented by $\boldsymbol{\omega} = V_1^{(1)}(\zeta)\, \mathbf{e}_\varphi$, where $V_1^{(1)}(\zeta)$ is a 1-harmonic scalar vorticity function, i.e., $\Delta_1 V_1^{(1)} = 0$.

The functions $\wp$ and $\boldsymbol{\omega}$ are related harmonic potentials (see Example 1). Consequently, $U_1^{(0)}(\zeta) = \wp(\zeta)/\mu$ and $V_1^{(1)}(\zeta)$ satisfy (3) for $k = 0$ and form the 0-harmonically analytic function $G_1^{(0)}(\zeta) = U_1^{(0)}(\zeta) + i\, V_1^{(1)}(\zeta)$, which under the condition $U_1^{(0)} \to 0$ and $V_1^{(1)} \to 0$ at $|\zeta| \to \infty$, is uniquely determined; see [30, Proposition 1]. Using the system (3) for $k = 0$, we can restate (38) as

$$\begin{aligned} \Delta_1 u_r &= \frac{\partial}{\partial r} U_1^{(0)} = \frac{\partial}{\partial z} V_1^{(1)}, \\ \Delta_0 u_z &= \frac{\partial}{\partial z} U_1^{(0)} = -\left(\frac{\partial}{\partial r} + \frac{1}{r}\right) V_1^{(1)}. \end{aligned} \tag{40}$$

With the identities

$$\begin{aligned} \Delta_1 \left(r\, U_1^{(0)}\right) &= 2\frac{\partial}{\partial r} U_1^{(0)}, \qquad & \Delta_1 \left(z\, V_1^{(1)}\right) &= 2\frac{\partial}{\partial z} V_1^{(1)}, \\ \Delta_0 \left(z\, U_1^{(0)}\right) &= 2\frac{\partial}{\partial z} U_1^{(0)}, \qquad & \Delta_0 \left(r\, V_1^{(1)}\right) &= 2\left(\frac{\partial}{\partial r} + \frac{1}{r}\right) V_1^{(1)}, \end{aligned}$$

equations (40) are integrated, and the components u_r and u_z can be represented in the form

$$\begin{aligned} u_r(\zeta) &= a\, r\, U_1^{(0)}(\zeta) + b\, z\, V_1^{(1)}(\zeta) + V_2^{(1)}(\zeta), \\ u_z(\zeta) &= c\, z\, U_1^{(0)}(\zeta) + d\, r\, V_1^{(1)}(\zeta) + U_2^{(0)}(\zeta), \end{aligned} \tag{41}$$

where a, b, c, and d are real-valued constants, and $U_2^{(0)}$ and $V_2^{(1)}$ are arbitrary 0-harmonic and 1-harmonic functions, respectively, i.e.,

$$\Delta_0 U_2^{(0)} = 0, \qquad \Delta_1 V_2^{(1)} = 0. \tag{42}$$

Substituting (41) into (40), we have

$$2a + 2b = 1, \qquad 2c - 2d = 1. \tag{43}$$

Then, substituting (41) into (39) and using (3) for $k = 0$, we obtain

$$\left(\frac{\partial}{\partial r} + \frac{1}{r}\right) V_2^{(1)} + \frac{\partial}{\partial z} U_2^{(0)} = 0, \tag{44}$$

provided that

$$a + d = 0, \qquad c - b = 0, \qquad 2a + c = 0. \tag{45}$$

Consequently, four constants a, b, c, and d satisfy five equations (43) and (45). However, these equations are dependent. Indeed, adding $a+b = 1/2$ and $c-b = 0$ and then subtracting $a + d = 0$ from the sum, we obtain the second equation in (43), i.e., $c - d = 1/2$. Excluding $c - d = 1/2$ from (43) and (45), we obtain four independent equations, whose unique solution is given by $a = -1/2$, $b = 1$, $c = 1$, and $d = 1/2$. Consequently, the representation (41) takes the form

$$\begin{aligned} u_r(\zeta) &= z\, V_1^{(1)}(\zeta) - \frac{1}{2}\, r\, U_1^{(0)}(\zeta) + V_2^{(1)}(\zeta), \\ u_z(\zeta) &= z\, U_1^{(0)}(\zeta) + \frac{1}{2}\, r\, V_1^{(1)}(\zeta) + U_2^{(0)}(\zeta). \end{aligned} \tag{46}$$

It follows from (42) and (44) that $U_2^{(0)}$ and $V_2^{(1)}$ form the 0-harmonically analytic function $G_2^{(0)}(\zeta) = U_2^{(0)}(\zeta) + i\, V_2^{(1)}(\zeta)$. Since $U_2^{(0)}$ and $V_2^{(1)}$ vanish at $|\zeta| \to \infty$, the function $G_2^{(0)}(\zeta)$ is uniquely determined; see [30, Proposition 1].

Thus, multiplying the first equation in (46) by i and adding with the second one in (46), we obtain the representation (36). $\square$

Remark 4. With $\mathbf{u}$ and $\wp$ vanishing at infinity in an unbounded region, (37) implies that $G_1^{(0)}|_\infty = 0$. According to Remark 3, such $G_1^{(0)}$ behaves as $\mathcal{O}\left(|\zeta|^{-2}\right)$ when $\zeta \to \infty$. Consequently, with this fact and $\mathbf{u}|_\infty = 0$, (36) implies that $G_2^{(0)}|_\infty = 0$.

Remark 5 (multiply connected region). The functions $G_1^{(0)}$ and $G_2^{(0)}$ are continuous and single-valued in multiply connected $\mathcal{D}_0^-$. Indeed, since the pressure and vorticity are continuous in $\mathcal{D}_0^-$, (37) implies that $G_1^{(0)}$ is continuous, and thus single-valued in $\mathcal{D}_0^-$. Also, since $\mathbf{u}$ is continuous in $\mathcal{D}_0^-$, the solution form (36) and continuity of $G_1^{(0)}$ imply that $G_2^{(0)}$ is continuous and thus single-valued in $\mathcal{D}_0^-$ as well.

Remark 6. The solution form (36) can be used for representing $\mathbf{u}$ for an inner Stokes flow problem. However, in this case, there is no requirement on $G_1^{(0)}$ and $G_2^{(0)}$ to vanish at infinity, and consequently $G_1^{(0)}$ and $G_2^{(0)}$ may not be uniquely determined.

The next proposition presents yet another solution form for the velocity field of axially symmetric Stokes flows in terms of 0-harmonically analytic and 1-harmonically analytic functions.

PROPOSITION 8. *Let $G_1^{(1)}(\zeta) = U_1^{(1)}(\zeta) + i\, V_1^{(2)}(\zeta)$ and $G_2^{(0)}(\zeta) = U_2^{(0)}(\zeta) + i\, V_2^{(1)}(\zeta)$ be 1-harmonically analytic and 0-harmonically analytic functions, respectively, vanishing at $|\zeta| \to \infty$; then the velocity field for an outer region in the axially symmetric case of the Stokes equations* (2) *can be represented in the form*

$$u_z(\zeta) + i\, u_r(\zeta) = r\, G_1^{(1)}(\zeta) + G_2^{(0)}(\zeta), \tag{47}$$

and the vorticity is determined by

$$\operatorname{curl} \mathbf{u} = -2 \operatorname{Re} G_1^{(1)}(\zeta)\, \mathbf{e}_\varphi.$$

Proof. The proof is similar to that of Proposition 7. $\square$

Let $\mathcal{D}^+$ and $\mathcal{D}^-$ denote the inner and outer regions with respect to the cross-section of the finite body of revolution in the rz-plane,[10] and let ℓ be the common boundary of $\mathcal{D}^+$ and $\mathcal{D}^-$. As in section 2.2, ℓ_+ denotes the right part of ℓ (i.e., ℓ with $r \geq 0$), which, being the contour of the body in the right half of the rz-plane, is either a closed curve or an open curve with the endpoints lying on the z-axis. The contour of the body in the rz-plane is, thus, $\ell = \ell_+ \bigcup \ell_-$, where ℓ_- is the reflection

[10] We always assume that the z-axis is the body's axis of revolution.

of ℓ_+ with respect to the z-axis. Also, $\mathcal{D}_0^+$ and $\mathcal{D}_0^-$ denote $\mathcal{D}^+$ with $r \geq 0$ and $\mathcal{D}^-$ $r \geq 0$, respectively, i.e., the right parts of the corresponding regions.

The representation (36) reduces an axially symmetric Stokes flow problem to a boundary-value problem for determining two 0-harmonically analytic functions.

Problem I (boundary-value problem for two 0-harmonically analytic functions). *Given a complex-valued function $f(\zeta)$ on ℓ_+ such that $f\left(-\overline{\zeta}\right) = \overline{f(\zeta)}$, find two* 0-*harmonically analytic functions $G_1^{(0)}(\zeta)$ and $G_2^{(0)}(\zeta)$ in (multiply connected) $\mathcal{D}_0^-$ that vanish at infinity and satisfy*

$$\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta) = f(\zeta), \quad \zeta \in \ell_+. \tag{48}$$

Remark 7. The relationship (48) is equivalent to $\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta) = f(\zeta)$ on ℓ. Indeed, with the symmetry condition for $f(\zeta)$ and $G^{(k)}(\zeta)$ (see (7)), $\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta) = f(\zeta)$ for $\zeta \in \ell_-$ can be restated as $\left(z + \frac{i}{2}\,r\right) G_1^{(0)}\left(-\overline{\zeta}\right) + G_2^{(0)}\left(-\overline{\zeta}\right) = f\left(-\overline{\zeta}\right)$ for $\zeta \in \ell_+$, or $\overline{\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta)} = \overline{f(\zeta)}$ for $\zeta \in \ell_+$, which is equivalent to (48).

There are two approaches to solving (48): (i) representing the functions $G_1^{(0)}(\zeta)$ and $G_2^{(0)}(\zeta)$ in the form of integrals or series in curvilinear coordinates associated with the geometry of the boundary ℓ_+ and finding unknown integral densities or series coefficients from (48); and (ii) reducing (48) to an integral equation based on Cauchy's integral formula for 0-harmonically analytic functions. For the second approach, Remark 7 is critical.

Problem I has a unique solution if it has only a zero homogeneous solution. The next proposition considers the homogeneous problem (48) for the outer region $\mathcal{D}_0^-$ and the inner region $\mathcal{D}_0^+$.

Proposition 9 (homogeneous boundary-value problems).

(i) *If functions $G_1^{(0)}(\zeta)$ and $G_2^{(0)}(\zeta)$ are* 0-*harmonically analytic in the outer (multiply connected) region $\mathcal{D}_0^-$ and vanish at infinity, then the homogeneous boundary-value problem*

$$\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta) = 0, \quad \zeta \in \ell_+, \tag{49}$$

has only a zero solution.

(ii) *If functions $G_1^{(0)}(\zeta)$ and $G_2^{(0)}(\zeta)$ are* 0-*harmonically analytic in the inner (multiply connected) region $\mathcal{D}_0^+$, and $\mathcal{D}_0^+$ consists of disjoint simply connected subregions $\mathcal{D}_j^+$, $1 \leq j \leq m$, then a solution to* (49) *is given by $G_1^{(0)}(\zeta) = a_j$ and $G_2^{(0)}(\zeta) = -a_j\left(z - \frac{i}{2}\,r\right)$ for $\zeta \in \mathcal{D}_j^+$, $1 \leq j \leq m$, where a_j is a real-valued constant.*

Proof. We first prove statement (i). Let a function $\Phi(\zeta)$ be defined by

$$\Phi(\zeta) = r\,G_1^{(0)}(\zeta)\left(\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta)\right), \tag{50}$$

and let $\mathcal{D}_R$ be a region in the right half of the rz-plane with piecewise smooth, *negatively* oriented boundary L, which consists of the positively oriented curve ℓ_+, segments of the z-axis, and semicircle L_R with large radius R (see Figure 2(a)). For multiply connected $\mathcal{D}_R$, the boundary L also contains crosscuts making $\mathcal{D}_R$ simply connected (see Figure 2(b)). Since $G_1^{(0)}$ and $G_2^{(0)}$ vanish at infinity as $\mathcal{O}\left(|\zeta|^{-2}\right)$ (see Remark 3), the integral of Φ over L_R vanishes at $R \to \infty$. Also, the function Φ vanishes on the z-axis. Consequently, we have

$$\lim_{R\to\infty} \oint_L \Phi\, d\zeta = \int_{\ell_+} \Phi\, d\zeta, \tag{51}$$

where, in the left-hand side, the integral over corresponding banks of the crosscuts vanishes, since Φ, being a combination of continuous single-valued functions $G_1^{(0)}$ and $G_2^{(0)}$ in $\mathcal{D}_0^-$ (see Remark 5), is continuous in $\mathcal{D}_0^-$ as well.

From (50) and Proposition 4(ii), it follows that

$$\operatorname{Im} \frac{\partial \Phi}{\partial \overline{\zeta}} = -\frac{1}{2} r \left(\operatorname{Im} G_1^{(0)}\right)^2 .$$

Consequently, using Proposition 5, we obtain

$$\lim_{R\to\infty} \operatorname{Re}\left[\oint_L \Phi\, d\zeta\right] = -\operatorname{Re}\left[2i \iint_{\mathcal{D}_0^-} \frac{\partial \Phi}{\partial \overline{\zeta}}\, drdz\right] = -\iint_{\mathcal{D}_0^-} r \left(\operatorname{Im} G_1^{(0)}\right)^2 drdz. \tag{52}$$

However, since $\Phi = 0$ on ℓ_+, (51) and (52) imply that the integral in the right-hand side in (52) vanishes, whence it follows that $\operatorname{Im} G_1^{(0)} = 0$ in $\mathcal{D}_0^-$. In this case, $\operatorname{Re} G_1^{(0)}$ can be a constant, which, however, equals zero, since $G_1^{(0)}$ vanishes at infinity. Thus, $G_1^{(0)} \equiv 0$ in $\mathcal{D}_0^-$, and the proof is finished.

Statement (ii) is proved similarly. Let $\widehat{\ell}_j$ be the part of ℓ_+ that corresponds to $\mathcal{D}_j^+$, $1 \le j \le m$. In this case, it is sufficient to conduct the proof for $\mathcal{D}_j^+$. The boundary of $\mathcal{D}_j^+$ is the closed, piecewise smooth, positively oriented curve L_j which either is the curve $\widehat{\ell}_j$ if $\widehat{\ell}_j$ is closed or consists of the curve $\widehat{\ell}_j$ and the segment of the z-axis connecting the endpoints of $\widehat{\ell}_j$ if $\widehat{\ell}_j$ is an open curve with the endpoints lying on the z-axis. For the same function (50), we have $\oint_{L_j} \Phi\, d\zeta = \int_{\widehat{\ell}_j} \Phi\, d\zeta$ and obtain a relationship similar to (52): $\operatorname{Re}[\oint_{L_j} \Phi\, d\zeta] = \iint_{\mathcal{D}_j^+} r(\operatorname{Im} G_1^{(0)})^2 drdz$. Consequently, since $\Phi = 0$ on $\widehat{\ell}_j$, we conclude that $G_1^{(0)}$ is a real-valued constant on each $\mathcal{D}_j^+$. Since $z - \frac{i}{2} r$ is a 0-harmonically analytic function in $\mathcal{D}_0^+$, statement (ii) follows. □

Alternatively, Proposition 9(i) can be proved based on the fact that the outer Stokes flow problem with the zero boundary condition $\mathbf{u}|_S = 0$ and with the velocity field and pressure vanishing at infinity has only a zero solution, i.e., $\mathbf{u} \equiv 0$; see [2, sections 2.8, 4.9]. Indeed, with (36), this implies that

$$\left(z - \frac{i}{2} r\right) G_1^{(0)}(\zeta) + G_2^{(0)}(\zeta) \equiv 0, \quad \zeta \in \mathcal{D}_0^-, \tag{53}$$

whence it follows that $\left(z - \frac{i}{2} r\right) G_1^{(0)}(\zeta)$ should be a 0-harmonically analytic function in $\mathcal{D}_0^-$. With (6), this requirement is equivalent to the equation

$$\frac{\partial}{\partial \overline{\zeta}} \left(\left(z - \frac{i}{2} r\right) G_1^{(0)}\right) = \frac{1}{4r} \left(\overline{\left(z - \frac{i}{2} r\right) G_1^{(0)}} - \left(z - \frac{i}{2} r\right) G_1^{(0)}\right), \quad \zeta \in \mathcal{D}_0^-,$$

which reduces to having

$$G_1^{(0)} = \overline{G_1^{(0)}}, \quad \zeta \in \mathcal{D}_0^- .$$

Consequently, $\operatorname{Im} G_1^{(0)} \equiv 0$ in $\mathcal{D}_0^-$, and (3) implies that $G_1^{(0)}$ is a real-valued constant a. Note that $z - \frac{i}{2}\, r$ is the 0-harmonically analytic function. Thus, the only solution to (53) in the class of 0-harmonically analytic functions is $G_1^{(0)} \equiv a$ and $G_2^{(0)} \equiv -a\left(z - \frac{i}{2}\, r\right)$. However, since $G_1^{(0)}$ and $G_2^{(0)}$ vanish at infinity, we conclude that $a = 0$, and statement (i) of Proposition 9 follows.

As one of the approaches to solving Problem I, we reduce (48) to an integral equation based on Cauchy's integral formula for 0-harmonically analytic functions.

Theorem 10 (integral equation in the axially symmetric case). *For the outer multiply connected region $\mathcal{D}_0^-$, Problem* I *reduces to the integral equation of the first kind,*

(54)

$$\frac{1}{\pi i}\int_{\ell_+}\left(\left[\left(z_1-\frac{i}{2}r_1\right)-\left(z-\frac{i}{2}r\right)\right]G_1^{(0)}(\tau)\,\Omega_+^{(0)}(\zeta,\tau)\,\frac{d\tau}{\tau-\zeta}\right.$$
$$\left. - \left[\left(z_1+\frac{i}{2}r_1\right)-\left(z-\frac{i}{2}r\right)\right]\overline{G_1^{(0)}(\tau)}\,\Omega_-^{(0)}(\zeta,\tau)\,\frac{d\overline{\tau}}{\overline{\tau}+\zeta}\right) = F(\zeta), \quad \zeta\in\ell_+,$$

where $\zeta = r + i\, z$ and $\tau = r_1 + i\, z_1$, and

$$F(\zeta) = f(\zeta) + \frac{1}{\pi i}\oint_{\ell} f(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\, d\tau. \tag{55}$$

If $\mathcal{D}_0^+$ consists of several disjoint, simply connected subregions $\mathcal{D}_j^+$, $1 \le j \le m$, and $\widehat{\ell}_j$ is the part of ℓ_+ that corresponds to $\mathcal{D}_j^+$ (obviously, $\ell_+ = \bigcup_{j=1}^m \widehat{\ell}_j$), then a solution to (54) *is determined with the accuracy of a real-valued constant a_j on each $\widehat{\ell}_j$.*[11] *Given a solution $\widetilde{G}_1^{(0)}(\zeta)$ to* (54)*, the constants are determined by*

$$a_j = \frac{1}{2}\widetilde{G}_1^{(0)}(\zeta) + \frac{1}{2\pi i}\oint_{\ell}\widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\, d\tau, \qquad \zeta\in\widehat{\ell}_j, \quad 1\le j\le m$$

(the right-hand side is constant for any $\zeta\in\widehat{\ell}_j$), and the solution to Problem I *takes the form*

$$G_1^{(0)}(\zeta) = \widetilde{G}_1^{(0)}(\zeta) - a_j, \qquad \zeta\in\widehat{\ell}_j, \quad 1\le j\le m. \tag{56}$$

Proof. The derivation of the integral equation (54) follows Muskhelishvili's approach, developed for solving 2D problems of an elastic medium; see [17, 18].

In view of Remark 7, Problem I for $\mathcal{D}_0^-$ is equivalent to the one for $\mathcal{D}^-$. Necessary and sufficient conditions for the functions $G_1^{(0)}$ and $G_2^{(0)}$ to be 0-harmonically analytic in the outer (multiply connected) region $\mathcal{D}^-$ and vanishing at infinity follow from the generalized Sokhotski–Plemelj formulae[12] and can be written in the form

$$G_1^{(0)}(\zeta) + \frac{1}{\pi i}\oint_{\ell} G_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\, d\tau = 0, \qquad \zeta\in\ell, \tag{57a}$$

$$G_2^{(0)}(\zeta) + \frac{1}{\pi i}\oint_{\ell} G_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\, d\tau = 0, \qquad \zeta\in\ell. \tag{57b}$$

[11] In other words, any function which is a real-valued constant on each $\widehat{\ell}_j$ is a homogeneous solution to (54).

[12] These formulae can be derived similarly to the Sokhotski–Plemelj formulae for ordinary analytic functions; see [1, formula (31.13)] and [9, formula (4.8)].

According to Remark 7, the boundary condition (48) on ℓ_+ is equivalent to (48) on ℓ. Expressing the boundary value of $G_2^{(0)}$ from (48),

$$G_2^{(0)}(\zeta) = f(\zeta) - \left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta), \quad \zeta \in \ell,$$

and substituting it into (57b), we obtain

$$\left(z - \frac{i}{2}\,r\right) G_1^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell \left(z_1 - \frac{i}{2}\,r_1\right) G_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = F(\zeta), \qquad \zeta \in \ell, \tag{58}$$

where $F(\zeta)$ is defined by (55). Then the combination (58) $-$ $\left(z - \frac{i}{2}\,r\right)\cdot$ (57a) reduces to the integral equation of the first kind

$$\frac{1}{\pi i}\oint_\ell \left[\left(z_1 - \frac{i}{2} r_1\right) - \left(z - \frac{i}{2} r\right)\right] G_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = F(\zeta), \qquad \zeta \in \ell, \tag{59}$$

which can equivalently be rewritten as (54).

Now we need to show that any function which is a real-valued constant on each $\widehat{\ell}_j$ is a homogeneous solution to (54) and that the solution to Problem I is determined by (56).

Let $\widetilde{G}_1^{(0)}$ solve (59). Adding the term $\left(z - \frac{i}{2}\,r\right)\widetilde{G}_1^{(0)}(\zeta)$ to the left-hand and right-hand sides of (59) and denoting $\widetilde{G}_2^{(0)}(\zeta) = f(\zeta) - \left(z - \frac{i}{2}\,r\right)\widetilde{G}_1^{(0)}(\zeta)$ for $\zeta \in \ell$, we rewrite (59) in the form

$$\begin{aligned}\left(z - \frac{i}{2}\,r\right)&\left(\widetilde{G}_1^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau\right)\\ &+ \widetilde{G}_2^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = 0, \qquad \zeta \in \ell.\end{aligned} \tag{60}$$

Let $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$ be determined by the generalized *Cauchy-type* integrals for the region $\mathcal{D}^+$ excluding its boundary ℓ:

$$\begin{aligned}\Phi^+(\zeta) &= \frac{1}{2\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \operatorname{int}\mathcal{D}^+,\\ \Psi^+(\zeta) &= \frac{1}{2\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \operatorname{int}\mathcal{D}^+.\end{aligned}$$

These functions are 0-harmonically analytic in $\operatorname{int}\mathcal{D}^+$, since $\mathcal{W}^{(0)}(\zeta,\tau)$ satisfies (6) for $k = 0$ with respect to ζ.[13] Then when ζ approaches ℓ from within $\mathcal{D}^+$, the boundary values of $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$ on ℓ are determined by the corresponding generalized Sokhotski–Plemelj formula:

$$\Phi^+(\zeta) = \frac{1}{2}\widetilde{G}_1^{(0)}(\zeta) + \frac{1}{2\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \qquad \zeta \in \ell, \tag{61a}$$

$$\Psi^+(\zeta) = \frac{1}{2}\widetilde{G}_2^{(0)}(\zeta) + \frac{1}{2\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \qquad \zeta \in \ell. \tag{61b}$$

[13] In fact, the generalized Cauchy-type integrals determine 0-harmonically analytic functions in the whole rz-plane, which are, however, discontinuous on the boundary ℓ; see [1, 9].

With (61a) and (61b), (60) reduces to

$$\left(z-\frac{i}{2}\,r\right)\Phi^{+}(\zeta)+\Psi^{+}(\zeta)=0,\qquad \zeta\in\ell, \tag{62}$$

which is a homogeneous boundary-value problem for the 0-harmonically analytic functions $\Phi^{+}(\zeta)$ and $\Psi^{+}(\zeta)$ in $\mathcal{D}^{+}$. Note that since $\Phi^{+}(\zeta)$ and $\Psi^{+}(\zeta)$ are determined by the generalized Cauchy-type integrals, they satisfy the symmetry condition (7) and thus, in view of Remark 7, the problem (62) is equivalent to (49) for $\mathcal{D}_0^{+}$. However, according to Proposition 9(ii), the only solution to (49) for $\mathcal{D}_0^{+}$ is $\Phi^{+}(\zeta)\equiv a_j$ and $\Psi^{+}(\zeta)\equiv -a_j\left(z-\frac{i}{2}\,r\right)$, $\zeta\in\mathcal{D}_j^{+}$, $1\le j\le m$, where a_j is a real-valued constant. Consequently, (61a) and (61b) imply that $G_1^{(0)}(\zeta)=\widetilde{G}_1^{(0)}(\zeta)-a_j$ and $G_2^{(0)}(\zeta)=\widetilde{G}_2^{(0)}(\zeta)+a_j\left(z-\frac{i}{2}\,r\right)$, $\zeta\in\widehat{\ell}_j$, $1\le j\le m$, satisfy (57a) and (57b), respectively, and thus are the boundary values of 0-harmonically analytic functions in $\mathcal{D}_0^{-}$. □

Remark 8. The first term in the integrand of (54) is regular:

$$\lim_{\tau\to\zeta}\frac{\left(z_1-\frac{i}{2}r_1\right)-\left(z-\frac{i}{2}r\right)}{\tau-\zeta}=\frac{i}{4}\left(\mathrm{e}^{-2i\lim\limits_{\tau\to\zeta}\arg[\tau-\zeta]}-3\right).$$

This expression is obtained by setting $\tau=\zeta+\rho\,\mathrm{e}^{i\beta}$ and passing $\rho\to 0$. Also, if $\zeta=\zeta(t)$ is a parameterization of smooth ℓ_{+}, then $\lim_{\tau\to\zeta}\arg[\tau-\zeta]=\arg[\zeta'(t)]$. The second term in (54) has a logarithmic singularity at $\tau=\zeta$ because of the function $\Omega_{-}^{(0)}(\zeta,\tau)$.

Remark 9 (integral equation for nonsmooth bodies). In the case when the surface of the body, i.e., spindle or lens, is nonsmooth, the necessary and sufficient condition for a function $G^{(0)}$ to be 0-harmonically analytic in $\mathcal{D}^{-}$ and vanishing at infinity follows from modified generalized Sokhotski–Plemelj formulae (see [1, (31.13a), (31.13b)]) and takes the form

$$h(\zeta)\,G^{(0)}(\zeta)+\frac{1}{\pi i}\oint_{\ell}G^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau=0,\quad \zeta\in\ell, \tag{63}$$

where $h(\zeta)=2-\alpha(\zeta)/\pi$ for $\operatorname{Re}\zeta\neq 0$ and $h(\zeta)=1+\cos(\alpha(\zeta)/2)$ for $\operatorname{Re}\zeta=0$ (if ℓ intersects the z-axis), and $\alpha(\zeta)$ is the angle between the right and left tangent vectors to the curve ℓ at the point ζ. For all points at which ℓ is smooth, $\alpha(\zeta)=\pi$ and $h(\zeta)=1$; see [1, (31.13a), (31.13b)].

As in the proof of Theorem 10, it can be shown that for bodies with nonsmooth surface, the integral equation (54) will be the same except for the right-hand side $F(\zeta)$, which in this case takes the form

$$F(\zeta)=h(\zeta)\,f(\zeta)+\frac{1}{\pi i}\oint_{\ell}f(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau. \tag{64}$$

Note that if $f(\zeta)$ is a 0-harmonically analytic function in $\mathcal{D}^{+}$, then $F(\zeta)=2f(\zeta)$.

The problem of axially symmetric Stokes flows that attracted much of the attention is the steady axially symmetric translation of a solid body of revolution with constant velocity v_z in the fluid. Let the body's axis of revolution coincide with the z-axis in the cylindrical coordinates. The fluid velocity $\mathbf{u}$ satisfies the Stokes equations (2) and no-slip boundary conditions on the surface S of the body:

$$\mathbf{u}|_{S}=v_z\,\mathbf{k}. \tag{65}$$

Also, $\mathbf{u}$ and $\wp$ vanish at infinity:

$$\mathbf{u}\big|_{\infty} = 0, \qquad \wp\big|_{\infty} = 0. \tag{66}$$

In terms of 0-harmonically analytic functions, this problem is a particular case of Problem I with $f(\zeta) = v_z$. Obviously, v_z satisfies the symmetry condition $f\left(-\overline{\zeta}\right) = \overline{f(\zeta)}$. In this case, since v_z is a 0-harmonically analytic function in $\mathcal{D}^+$, the right-hand side of (54) reduces to $F(\zeta) = 2v_z$.

The next proposition shows that the resisting (drag) force, exerted on the body of revolution in the axially symmetric translation, can be expressed in terms of $G_1^{(0)}$.

Proposition 11 (drag force in axially symmetric translation). *For the axially symmetric Stokes flow problem for the body translating along the z-axis,*[14] *let the velocity field be represented by* (36) *with the boundary conditions* (65) *and* (66). *The resisting force, exerted on the body, can be represented in two equivalent forms*

$$F_z = 2\pi\mu \, \mathrm{Re}\left[\int_{\ell_+} r\, G_1^{(0)}(\zeta)\, d\zeta\right], \tag{67a}$$

$$F_z = -4\pi\mu \lim_{z\to\infty} \left(z^2 \, \mathrm{Re}\, G_1^{(0)}(\zeta)\Big|_{r=0}\right), \tag{67b}$$

where ℓ_+ in (67a) *is positively oriented with respect to $\mathcal{D}_0^+$.*

Proof. First, we prove (67a).

The resulting force exerted by the fluid on a solid body is defined as the integral over the body's surface S:

$$\mathbf{F} = \iint_S \mathbf{P}_n \, dS, \qquad \mathbf{P}_n = 2\mu \frac{\partial \mathbf{u}}{\partial n} + \mu\left[\mathfrak{n} \times \operatorname{curl} \mathbf{u}\right] - \wp\, \mathfrak{n}, \tag{68}$$

where $\mathfrak{n} = n_r\, \mathbf{e}_r + n_z\, \mathbf{k}$ is the outer normal to the body's surface with $n_r = \frac{\partial r}{\partial n}$ and $n_z = \frac{\partial z}{\partial n}$; see [11].

It can be shown that for the axially symmetric velocity field with the boundary conditions (65),

$$\frac{\partial \mathbf{u}}{\partial n} = -[\mathfrak{n} \times \operatorname{curl} \mathbf{u}] \quad \text{on } S. \tag{69}$$

Indeed, in this case, in the cylindrical coordinates (r, φ, z), the velocity field is independent of the angular coordinate φ: $\mathbf{u} = u_r(r,z)\, \mathbf{e}_r + u_z(r,z)\, \mathbf{k}$ and $u_\varphi \equiv 0$, and we have $\operatorname{curl} \mathbf{u} = \left(\frac{\partial u_r}{\partial z} - \frac{\partial u_z}{\partial r}\right) \mathbf{e}_\varphi$. With $\mathfrak{n} = \frac{\partial r}{\partial n}\, \mathbf{e}_r + \frac{\partial z}{\partial n}\, \mathbf{k}$, we obtain

$$-[\mathfrak{n} \times \operatorname{curl} \mathbf{u}] = \frac{\partial z}{\partial n}\left(\frac{\partial u_r}{\partial z} - \frac{\partial u_z}{\partial r}\right) \mathbf{e}_r + \frac{\partial r}{\partial n}\left(\frac{\partial u_z}{\partial r} - \frac{\partial u_r}{\partial z}\right) \mathbf{k}. \tag{70}$$

Let (s, φ, n) be a characteristic coordinate system with the right-handed orthogonal basis $(\mathfrak{s}, \mathbf{e}_\varphi, \mathfrak{n})$, in which s has *negative* orientation. Then using the relationships

$$\frac{\partial r}{\partial s} = \frac{\partial z}{\partial n}, \qquad \frac{\partial r}{\partial n} = -\frac{\partial z}{\partial s}$$

[14] The z-axis is the body's axis of revolution.

and the continuity equation (39), we have

$$\begin{aligned}\frac{\partial z}{\partial n}\left(\frac{\partial u_r}{\partial z}-\frac{\partial u_z}{\partial r}\right)&=\frac{\partial u_r}{\partial n}-\frac{\partial u_z}{\partial s}-\frac{u_r}{r}\frac{\partial z}{\partial s},\\ \frac{\partial r}{\partial n}\left(\frac{\partial u_z}{\partial r}-\frac{\partial u_r}{\partial z}\right)&=\frac{\partial u_z}{\partial n}+\frac{\partial u_r}{\partial s}+\frac{u_r}{r}\frac{\partial r}{\partial s}.\end{aligned}\tag{71}$$

The boundary conditions (65) imply that on the surface S, the first and second equations in (71) reduce to $\partial u_r/\partial n$ and $\partial u_z/\partial n$, respectively, and thus (69) holds.

Consequently, for the axially symmetric translation, $\mathbf{P}_n$ in (68) is determined on the surface S by

$$\mathbf{P}_n=-\mu\left[\mathfrak{n}\times\operatorname{curl}\mathbf{u}\right]-\wp\,\mathfrak{n}\quad\text{on } S.\tag{72}$$

Finally, in the cylindrical coordinates, we have $dS=r\,ds\,d\varphi$, where ds is the differential of the curve length. In the axially symmetric case, $\operatorname{curl}\mathbf{u}$ and $\wp$ are represented by (37), and $\mathbf{F}$ has the component in the direction $\mathbf{k}$ only. Thus, the corresponding component of the integral (68) with (72) takes the form

$$F_z=(\mathbf{F}\cdot\mathbf{k})=-2\pi\mu\int_{\ell_+} r\left((\mathfrak{n}\cdot\mathbf{e}_r)\,\operatorname{Im}G_1^{(0)}(\zeta)+(\mathfrak{n}\cdot\mathbf{k})\,\operatorname{Re}G_1^{(0)}(\zeta)\right)ds.$$

With $(\mathfrak{n}\cdot\mathbf{e}_r)\,ds=dz$ and $(\mathfrak{n}\cdot\mathbf{k})\,ds=-dr$ (for positively oriented ℓ_+), the formula (67a) follows.

The formula (67b) follows from (67a) and Proposition 6. □

Remark 10. The formula (67a) is invariant with respect to adding a real-valued constant to $G_1^{(0)}|_{\ell_+}$. This means that the drag force F_z can be calculated by (67a) for any solution of the integral equation (54).

4. Exact solutions to axially symmetric Stokes flow problems. As an illustration to the developed framework, this section presents exact solutions in the form (36) to the problem of Stokes flows due to the steady axially symmetric translation of a solid body of revolution. In this case, the boundary conditions for the Stokes equations (2) are given by (65) and (66).

Example 5 (axially symmetric translation of a solid sphere). Let (R,ϑ,φ) be the spherical coordinates related to the cylindrical coordinates in the ordinary way, and let a solid sphere be centered at the origin and have radius c. For the axially symmetric translation of the sphere, the 0-harmonically analytic functions $G_1^{(0)}$ and $G_2^{(0)}$ in (36) are determined in the region $R\geq c$ by

$$G_1^{(0)}(R,\vartheta)=\frac{3v_zc}{2}\,R^{-2}\,\mathrm{e}^{i\vartheta},\qquad G_2^{(0)}(R,\vartheta)=-\frac{v_zc^3}{8}\,R^{-3}\left(1+3\,\mathrm{e}^{2i\vartheta}\right),\tag{73}$$

and the drag force $F_z=-6\pi c\mu v_z$ follows from (67b).

Detail. Let the velocity field be determined by (36). Representing $G_1^{(0)}$ and $G_2^{(0)}$ for the region exterior to the sphere by (15) with $k=0$,

$$\begin{aligned}G_1^{(0)}(R,\vartheta)&=\sum_{n=1}^{\infty}A_n\,R^{-n-1}\left\{n\,\mathrm{P}_n(\cos\vartheta)-i\,\mathrm{P}_n^{(1)}(\cos\vartheta)\right\},\\ G_2^{(0)}(R,\vartheta)&=\sum_{n=1}^{\infty}B_n\,R^{-n-1}\left\{n\,\mathrm{P}_n(\cos\vartheta)-i\,\mathrm{P}_n^{(1)}(\cos\vartheta)\right\},\end{aligned}$$

and substituting these series into (48) with $f(\zeta) = v_z$, we obtain $2A_1/(3c) = v_z$ and the following system:

$$c^2 \frac{n(n-1)}{2(2n-1)} A_{n-1} + \frac{(n+1)(3n+4)}{2(2n+3)} A_{n+1} + n\, B_n = 0, \qquad n \geq 1, \tag{74a}$$

$$c^2 \frac{(n-1)}{2(2n-1)} A_{n-1} + \frac{(3n+4)}{2(2n+3)} A_{n+1} + B_n = 0, \qquad n \geq 1. \tag{74b}$$

The difference (74a) $-\, n \cdot$ (74b) reduces to $A_{n+1} = 0$ for $n \geq 1$, and, hence, it follows from either (74a) or (74b) that $B_1 = 0$, $B_2 = -A_1 c^2/6$, and $B_n = 0$ for $n \geq 3$.

The next two examples present analytical solutions for the axially symmetric translation of solid prolate and oblate spheroids in the form (36). Analytical solutions to these problems in terms of a stream function can be found in [11].

Example 6 (axially symmetric translation of a solid prolate spheroid). Let a solid prolate spheroid be described in the prolate spheroidal coordinates (19) by $\xi = \xi_0$. For the axially symmetric translation of the prolate spheroid, the 0-harmonically analytic functions $G_1^{(0)}$ and $G_2^{(0)}$ in (36) are determined in the region $\xi \geq \xi_0$ by

$$\begin{aligned} G_1^{(0)}(\xi,\eta) &= \frac{q}{c} \, \frac{(\cos\eta + i\, \coth\xi \sin\eta)}{\cosh^2\xi - \cos^2\eta}, \\ G_2^{(0)}(\xi,\eta) &= \frac{q}{2}\left(1 + \cosh^2\xi_0\right)\left(\ln\left(\coth[\xi/2]\right) - \frac{\sinh[2\xi] + i\, \sin[2\eta]}{2\sinh\xi\left(\cosh^2\xi - \cos^2\eta\right)}\right), \end{aligned} \tag{75}$$

where $q = 2v_z \big/ \big(\left(1 + \cosh^2\xi_0\right)\ln\left(\coth[\xi_0/2]\right) - \cosh\xi_0 \big)$. The drag force $F_z = -4\pi c\mu q$ follows from (67b) and (75).

Detail. For the region exterior to the prolate spheroid, the 0-harmonically analytic functions $G_1^{(0)}$ and $G_2^{(0)}$ are determined by (20) for $k = 0$:

$$\begin{aligned} G_1^{(0)}(\xi,\eta) &= \sum_{n=1}^{\infty} A_n \left\{ n(n+1)\, \mathrm{Q}_n(\cosh\xi)\mathrm{P}_n(\cos\eta) + i\, \mathrm{Q}_n^{(1)}(\cosh\xi)\mathrm{P}_n^{(1)}(\cos\eta) \right\}, \\ G_2^{(0)}(\xi,\eta) &= \sum_{n=1}^{\infty} B_n \left\{ n(n+1)\, \mathrm{Q}_n(\cosh\xi)\mathrm{P}_n(\cos\eta) + i\, \mathrm{Q}_n^{(1)}(\cosh\xi)\mathrm{P}_n^{(1)}(\cos\eta) \right\}, \end{aligned} \tag{76}$$

and the boundary-value problem $((z - \frac{i}{2}\, r)G_1^{(0)} + G_2^{(0)})|_{\xi=\xi_0} = v_z$ reduces to

$$L_n(\xi_0)\, A_{n-1} + M_n(\xi_0)\, A_{n+1} + \mathrm{Q}_n^{(1)}(\cosh\xi_0)\, B_n = 0, \qquad n \geq 1, \tag{77a}$$

$$K_0(\xi_0)\, A_1 = v_z, \tag{77b}$$

$$N_n(\xi_0)\, A_{n-1} + K_n(\xi_0)\, A_{n+1} + n(n+1)\mathrm{Q}_n(\cosh\xi_0)\, B_n = 0, \qquad n \geq 1, \tag{77c}$$

where the functions $L_n(\xi_0)$, $M_n(\xi_0)$, $N_n(\xi_0)$, and $K_n(\xi_0)$ are defined by

$$\begin{aligned}
L_n(\xi_0) &= c\,\frac{(n-1)}{2n-1}\left(\cosh\xi_0\,\mathrm{Q}^{(1)}_{n-1}(\cosh\xi_0) + \frac{n}{2}\sinh\xi_0\,\mathrm{Q}_{n-1}(\cosh\xi_0)\right),\\
M_n(\xi_0) &= c\,\frac{(n+2)}{2n+3}\left(\cosh\xi_0\,\mathrm{Q}^{(1)}_{n+1}(\cosh\xi_0) - \frac{(n+1)}{2}\sinh\xi_0\,\mathrm{Q}_{n+1}(\cosh\xi_0)\right),\\
N_n(\xi_0) &= c\,\frac{n(n-1)}{2n-1}\left(n\cosh\xi_0\,\mathrm{Q}_{n-1}(\cosh\xi_0) + \frac{1}{2}\sinh\xi_0\,\mathrm{Q}^{(1)}_{n-1}(\cosh\xi_0)\right),\\
K_n(\xi_0) &= c\,\frac{(n+1)(n+2)}{2n+3}\left((n+1)\cosh\xi_0\,\mathrm{Q}_{n+1}(\cosh\xi_0)\right.\\
&\qquad\qquad\left.-\frac{1}{2}\sinh\xi_0\,\mathrm{Q}^{(1)}_{n+1}(\cosh\xi_0)\right).
\end{aligned}$$

The combination $\mathrm{Q}^{(1)}_n(\cosh\xi_0)\cdot(77\text{c}) - n(n+1)\mathrm{Q}_n(\cosh\xi_0)\cdot(77\text{a})$ reduces to

$$\left(-\widetilde{A}_{n-1} + \widetilde{A}_{n+1}\right)\delta_n(\xi_0) = 0, \qquad n \geq 2, \tag{78}$$

where $\widetilde{A}_n = \frac{n(n+1)}{2(2n+1)}\,A_n$ and

$$\begin{aligned}
\delta_n(\xi_0) &= \left(1+\cosh^2\xi_0\right)\mathrm{Q}_n(\cosh\xi_0)\,\mathrm{Q}^{(1)}_n(\cosh\xi_0)\\
&\quad + \sinh\xi_0\cosh\xi_0\left[\left(\mathrm{Q}^{(1)}_n(\cosh\xi_0)\right)^2 - n(n+1)\left(\mathrm{Q}_n(\cosh\xi_0)\right)^2\right].
\end{aligned}$$

From (77b), we have $\widetilde{A}_1 = v_z/(3K_0(\xi_0)) = q/(2c)$, and solving (77a) and (77c) for $n=1$, we obtain $\widetilde{A}_2 = 0$ and $B_1 = 0$. Consequently, since $\delta_n(\xi_0) \neq 0$ for $n \geq 2$, it follows from (78) that $\widetilde{A}_{2m+1} = \widetilde{A}_1$, $\widetilde{A}_{2m} = 0$, $B_{2m-1} = 0$, and

$$B_{2m} = -c\left(1+\cosh^2\xi_0\right)\frac{(4m+1)}{2m(2m+1)}\,\widetilde{A}_{2m-1}, \qquad m \geq 1.$$

Thus, the series (76) take the form

$$\begin{aligned}
G^{(0)}_1(\xi,\eta) &= 2\widetilde{A}_1\sum_{m=0}^{\infty}(4m+3)\left\{\mathrm{Q}_{2m+1}(\cosh\xi)\,\mathrm{P}_{2m+1}(\cos\eta)\right.\\
&\qquad\left.+\frac{i}{(2m+1)(2m+2)}\,\mathrm{Q}^{(1)}_{2m+1}(\cosh\xi)\,\mathrm{P}^{(1)}_{2m+1}(\cos\eta)\right\},\\
G^{(0)}_2(\xi,\eta) &= -c\left(1+\cosh^2\xi_0\right)\widetilde{A}_1\sum_{m=1}^{\infty}(4m+1)\left\{\mathrm{Q}_{2m}(\cosh\xi)\,\mathrm{P}_{2m}(\cos\eta)\right.\\
&\qquad\left.+\frac{i}{2m(2m+1)}\,\mathrm{Q}^{(1)}_{2m}(\cosh\xi)\,\mathrm{P}^{(1)}_{2m}(\cos\eta)\right\},
\end{aligned}\tag{79}$$

which reduce to (75) with the representations (see [3])

$$\begin{aligned}
\frac{1}{\cosh\xi-\cos\eta} &= \sum_{n=0}^{\infty}(2n+1)\,\mathrm{Q}_n(\cosh\xi)\,\mathrm{P}_n(\cos\eta),\\
\frac{\sin\eta}{\sinh\xi\,(\cosh\xi-\cos\eta)} &= \sum_{n=1}^{\infty}\frac{(2n+1)}{n(n+1)}\,\mathrm{Q}^{(1)}_n(\cosh\xi)\,\mathrm{P}^{(1)}_n(\cos\eta)
\end{aligned}\tag{80}$$

and the fact that $\mathrm{P}_n^{(k)}(-\cos\eta) = (-1)^{n+k}\,\mathrm{P}_n^{(k)}(\cos\eta)$ for integers n and k.

Example 7 (axially symmetric translation of a solid oblate spheroid). Let a solid oblate spheroid be described in the oblate spheroidal coordinates (22) by $\xi = \xi_0$. For the axially symmetric translation of the oblate spheroid, the 0-harmonically analytic functions $G_1^{(0)}$ and $G_2^{(0)}$ in (36) are determined in the region $\xi \geq \xi_0$ by

$$\begin{aligned} G_1^{(0)}(\xi,\eta) &= -\frac{q}{c}\,\frac{(\cos\eta + i\,\tanh\xi\sin\eta)}{\sinh^2\xi + \cos^2\eta}, \\ G_2^{(0)}(\xi,\eta) &= \frac{q}{2}\left(\sinh^2\xi_0 - 1\right)\left(\operatorname{arccot}[\sinh\xi] - \frac{\sinh[2\xi] - i\,\sin[2\eta]}{2\cosh\xi\left(\sinh^2\xi + \cos^2\eta\right)}\right), \end{aligned} \tag{81}$$

where $q = -2v_z \big/ \left(\sinh\xi_0 - \left(\sinh^2\xi_0 - 1\right)\operatorname{arccot}[\sinh\xi_0]\right)$. The drag force $F_z = 4\pi c\mu q$ follows from (67b) and (81).

Detail. The solution (81) is obtained similarly to (76).

We also solve the integral equation (54) for the axially symmetric translation of the sphere and prolate and oblate spheroids and compare the obtained solutions with the corresponding closed-form analytical solutions (73), (76), and (81).

Remark 11. In Problem I, the boundary conditions (65), corresponding to the axially symmetric translation, satisfy $f\left(\overline{\zeta}\right) = \overline{f(\zeta)}$ on ℓ_+. In this case, if, for a given body, $\mathcal{D}_0^+$ is symmetric with respect to the r-axis, then we have the symmetry condition[15] $G_1^{(0)}\left(\overline{\zeta}\right) = -\overline{G_1^{(0)}(\zeta)}$. If also $\mathcal{D}_0^+$ is simply connected, then this condition implies that a solution to (54) is uniquely determined (see Theorem 10).

For the sphere and prolate and oblate spheroids, $G_1^{(0)}\left(\overline{\zeta}\right) = -\overline{G_1^{(0)}(\zeta)}$, and a solution to (54) is uniquely determined (see Remark 11). For example, for the sphere of unit radius, we parameterize ℓ_+ by $r(t) = \cos t$ and $z(t) = \sin t$, $t \in [-\pi/2, \pi/2]$, and represent the boundary-value of the function $G_1^{(0)}$ on ℓ_+ by

$$G_1^{(0)}(\zeta(t)) = \sum\nolimits_{k=1}^{n} \left(a_k \sin[k\,t] + i\,b_k\cos[(k-1)t]\right), \qquad t \in [-\pi/2, \pi/2], \tag{82}$$

where the real and imaginary parts are odd and even functions of t, respectively, and the unknown coefficients a_k and b_k can be found by various techniques, e.g., by minimizing the total quadratic error of (54) with (82). For the sphere, we can take only first two terms in (82), i.e., $n = 2$, to obtain an exact solution coinciding with (73). For prolate and oblate spheroids, we solve (54) using the same approach with corresponding parameterization of r and z and obtain that with only $n = 8$ in (82), the drag force F_z, being compared to the corresponding exact values, has the relative error of 10^{-5}.

The next two examples illustrate solutions to the integral equation (54) for bispheroids (two separate spheroids of equal size and having the same axis of revolution) and a torus of elliptical cross-section for various values of a geometrical parameter.[16] The Stokes flow problem for two spheres was solved analytically in [23, 26, 30]. The axially symmetric problem of sedimentation of bispheroids was considered in [25] (see also [11]); however, the pressure was not investigated in this case.

Example 8 (axially symmetric translation of solid bispheroids). Let the centers of bispheroids lie on the z-axis, which is the axis of revolution, and have coordinates

[15] This condition should not be confused with the symmetry condition (7).

[16] The integral equation (54) holds for multiply connected regions.

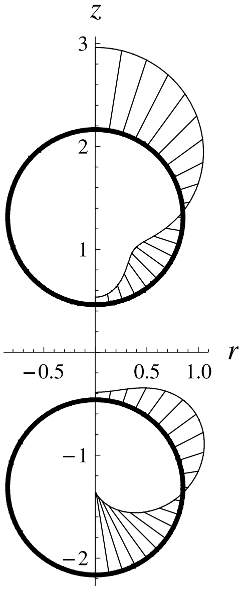
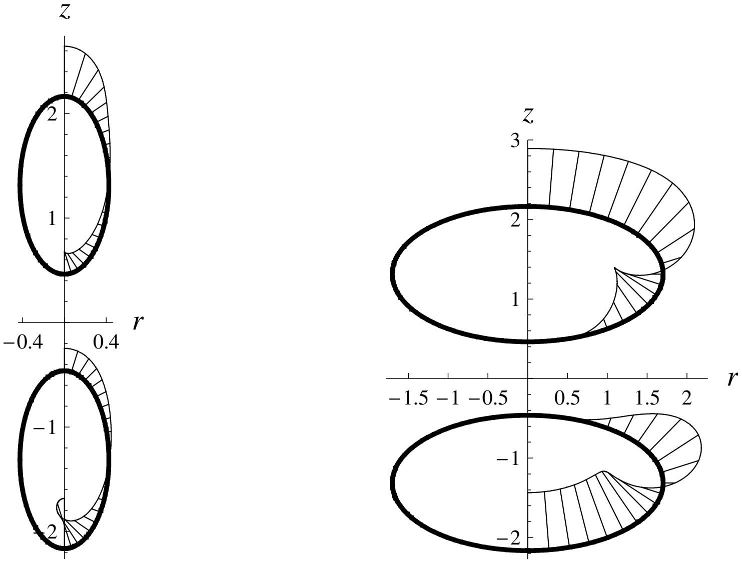


FIG. 3. *Profile of the pressure,* $\wp/(2v_z\mu)$*, on the surface of the solid bispheres (*$\varkappa = 1$*) in the axially symmetric translation along the* z*-axis.*



(a) $\wp/(8v_z\mu)$, prolate bi-spheroids

(b) $\wp/(v_z\mu)$, oblate bi-spheroids

FIG. 4. *Profile of the pressure on the surface of the solid bispheroids in the axially symmetric translation along the* z*-axis:* (a) $\wp/(8v_z\mu)$*, prolate bispheroids with* $\varkappa = 0.5$*;* (b) $\wp/(v_z\mu)$*, oblate bispheroids with* $\varkappa = 2$*.*

$z = \coth 1$ (upper spheroid) and $z = -\coth 1$ (lower spheroid). Let the upper spheroid be parameterized in the right half of the rz-plane by $r(t) = \varkappa \sin t/\sinh 1$, $z(t) = \coth 1 - \cos t/\sinh 1$, $t \in [0, \pi]$, where $\varkappa$ is a positive parameter. The case $\varkappa = 1$ corresponds to bispheres, for which a closed form solution to the axially symmetric

Stokes flow problem can be obtained in terms of a stream function [23, 30]. On ℓ_+ for $z \geq 0$ (upper spheroid), we represent $G_1^{(0)}(\zeta)$ by

$$G_1^{(0)}(\zeta(t)) = a_0 + \sum_{k=1}^{n} \left(a_k \cos[k\,t] + i\, b_k \cos[(k-1)t] \right), \qquad t \in [0, \pi],$$

and on ℓ_+ for $z \leq 0$ (lower spheroid), we determine $G_1^{(0)}(\zeta)$ using $G_1^{(0)}(\overline{\zeta}) = -\overline{G_1^{(0)}(\zeta)}$ (see Remark 11). Since for the bispheroids $\mathcal{D}_0^+$ is a doubly connected region, the function $g(\zeta)$ defined by $g(\zeta) = a_0$ on ℓ_+ for $z \geq 0$ and by $g(\zeta) = -a_0$ on ℓ_+ for $z \leq 0$ is a homogeneous solution to (54). Consequently, we solve (54) for $\widetilde{G}_1^{(0)}(\zeta(t)) = G_1^{(0)}(\zeta(t)) - a_0$ and then determine the constant a_0 by (see Theorem 10)

$$a_0 = -\frac{1}{2}\widetilde{G}_1^{(0)}(\zeta) - \frac{1}{2\pi i}\oint_{\ell} \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \qquad \zeta \in \ell_+.$$

Figures 3 and 4 illustrate profiles of the pressure on the surface of the bispheres ($\varkappa = 1$) and bispheroids for $\varkappa = 0.5$ and $\varkappa = 2$. Table 1 presents the ratio, d_z, of the drag, exerted on one of the two spheroids and calculated by (67a), to the drag of a single same-size spheroid[17] for $\varkappa = 0.5$, 1, and 2.

As another illustration, we solve the integral equation (54) for the axially symmetric translation of a torus of elliptical cross-section.[18] To the best of our knowledge, only the case of the torus of circular cross-section in this Stokes flow problem was addressed in the literature; see [19, 27, 10, 30].

Example 9 (axially symmetric translation of a solid torus of elliptical cross-section). Let the surface of a torus of elliptical cross-section be parameterized in the right half of the rz-plane by $r(t) = 2 + \cos t$, $z(t) = \varkappa \sin t$, $t \in [-\pi, \pi]$, where $\varkappa$ is a positive parameter. The case $\varkappa = 1$ corresponds to the torus of circular cross-section, for which a closed form solution can be obtained in terms of a stream function [10, 30]. According to Remark 11, $G_1^{(0)}\left(\overline{\zeta}\right) = -\overline{G_1^{(0)}(\zeta)}$, and a solution to (54) is uniquely determined and can be represented on ℓ_+ by

$$G_1^{(0)}(\zeta(t)) = \sum_{k=1}^{n} \left(a_k \sin[k\,t] + i\, b_k \cos[(k-1)t] \right), \qquad t \in [-\pi, \pi].$$

Figures 5, 6, and 7 illustrate profiles of the pressure on the surface of the torus of elliptical cross-section in the axially symmetric translation for $\varkappa = 1$, 0.5, and 2, respectively. Table 2 presents the drag F_z, calculated by (67a) and normalized to the drag of the circumscribed sphere with the radius 3 for $\varkappa = 0.5$, 1, and 2.

TABLE 1

The ratio, d_z, of the drag, exerted on one of the two spheroids to the drag of a single same-size spheroid in the axially symmetric translation along the z-axis for $\varkappa = 0.5$, 1, and 2.

$\varkappa$	0.5	1.0[§]	2
d_z	0.7736	0.7025[♭]	0.6332

[§] The case corresponds to the bispheres.

[♭] The value coincides with the one in [23]; see also [30].

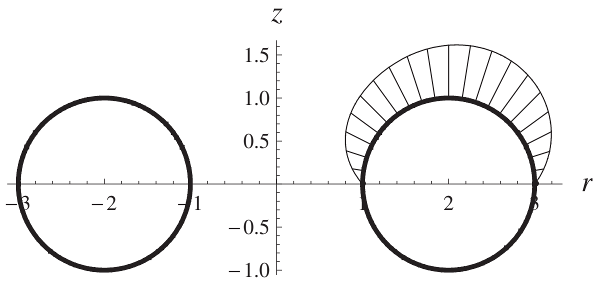


FIG. 5. *Profile of the pressure, $\wp/(v_z\mu)$, on the surface of the solid torus of circular cross-section ($\varkappa = 1$) in the axially symmetric translation along the z-axis (for $z < 0$, the profile is antisymmetric).*

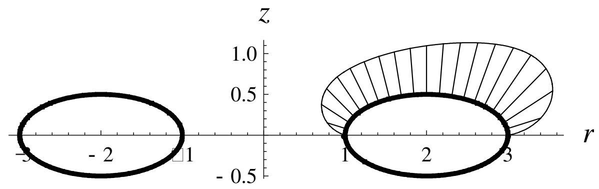


FIG. 6. *Profile of the pressure, $\wp/(v_z\mu)$, on the surface of the solid torus of elliptical cross-section for $\varkappa = 0.5$ in the axially symmetric translation along the z-axis (for $z < 0$, the profile is antisymmetric).*

The developed framework is not, however, limited to the case of smooth bodies.

Remark 9 states that for bodies with nonsmooth surfaces, we need to solve the integral equation (54) with the right-hand side $F(\zeta)$ in the form of (64). Since $f(\zeta) = v_z$ is a 0-harmonically analytic function in $\mathcal{D}^+$, (64), corresponding to the modified generalized Sokhotski–Plemelj formula, reduces to $F(\zeta) = 2v_z$. Thus, based on this fact and Remark 9, we conclude that (54) with the right-hand side of $2v_z$ holds for bodies with smooth and nonsmooth surfaces. In the next example, we solve (54) for a spindle-shaped body and biconvex lens and compare solutions to the analytical solutions obtained in our previous work [32, 31].

Example 10 (axially symmetric translation of a solid spindle and biconvex lens). In the first quadrant of the rz-plane, let the surface of the spindle be parameterized by

$$r(t) = \frac{\cos\left[\frac{2(\pi-\eta)t}{\pi}\right] + \cos\eta}{\sin\eta}, \quad z(t) = \frac{\sin\left[\frac{2(\pi-\eta)t}{\pi}\right]}{\sin\eta}, \quad t \in [0, \pi/2]\,,$$

and let the surface of the biconvex lens be parameterized by

$$r(t) = \frac{\sin\left[(\pi-\eta)\left(1-\frac{2t}{\pi}\right)\right]}{\sin\eta}, \;\; z(t) = \frac{\cos\left[(\pi-\eta)\left(1-\frac{2t}{\pi}\right)\right] + \cos\eta}{\sin\eta}, \;\; t \in [0, \pi/2]\,,$$

where $\eta \in (0, \pi)$[19] is a parameter coinciding with the coordinate η in the bispherical

[17] This spheroid has the size of one spheroid in the bispheroids.

[18] Special functions associated with the geometry of a torus of elliptical cross-section were considered in [15] and could potentially be used for obtaining analytical solutions to the corresponding Stokes flow problem.

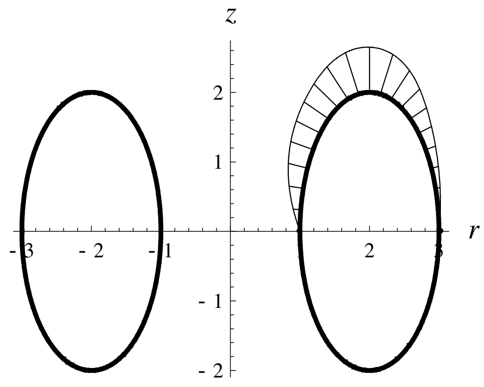


Fig. 7. *Profile of the pressure, $\wp/(v_z\mu)$, on the surface of the solid torus of elliptical cross-section for $\varkappa = 2$ in the axially symmetric translation along the z-axis (for $z < 0$, the profile is antisymmetric).*

Table 2

Normalized drag, $d_z = F_z/(18\pi\mu v_z)$, for the torus of elliptical cross-section in the axially symmetric translation along the z-axis for $\varkappa = 0.5$, 1, and 2.

$\varkappa$	0.5	1.0§	2
d_z	0.8667	0.9072♭	0.9982

§The case corresponds to the torus of circular cross-section.

♭The value coincides with the one in [10]; see also [30].

and toroidal coordinates (see [32, 31]).

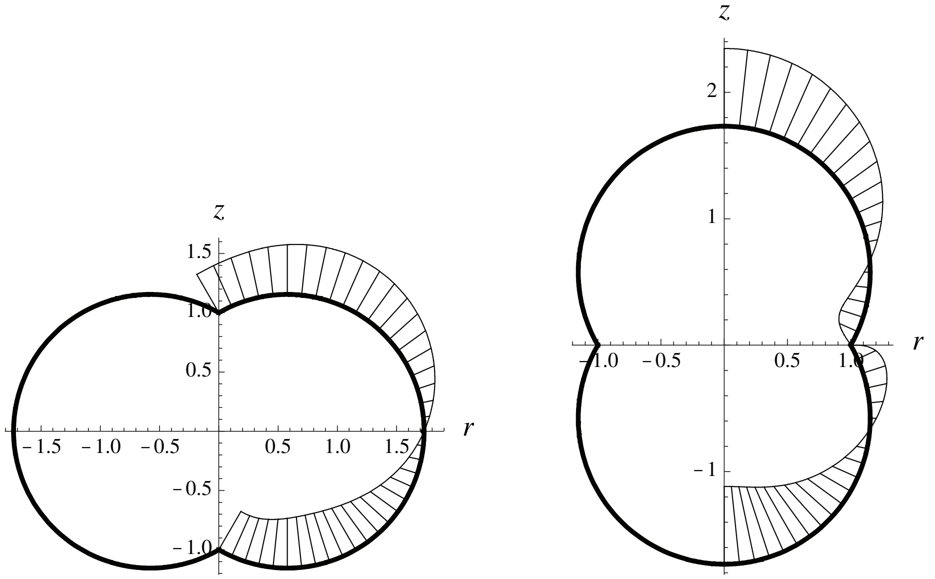
In both cases, the function $G_1^{(0)}$ can be represented on ℓ_+ for $z \geq 0$ by

$$G_1^{(0)}(\zeta(t)) = \sum_{k=1}^{n} (a_k + i\, b_k)\, T_{k-1}\left(\frac{4t}{\pi} - 1\right), \qquad t \in [0, \pi/2], \tag{83}$$

where $T_k(t)$ is the Chebyshev polynomial of the first kind and can be determined on ℓ_+ for $z < 0$ by the symmetry condition $G_1^{(0)}\left(\overline{\zeta}\right) = -\overline{G_1^{(0)}(\zeta)}$ (see Remark 11). It is known that both the pressure and vorticity are unbounded at $t = \pi/2$ for the spindle with $\eta > 2\pi/3$ and are unbounded at $t = 0$ for the biconvex lens with $\eta > \pi/2$ (see [32, 31]). Consequently, the representation (83) is valid for the spindle with $\eta < 2\pi/3$ and for the biconvex lens with $\eta \leq \pi/2$. Figure 8 shows profiles of the pressure on the surface of the solid spindle and biconvex lens with $\eta = \pi/3$ and $n = 16$. For the spindle and biconvex lens, the pressure, $\mu\,\mathrm{Re}\, G_1^{(0)}$, and vorticity, $\mathrm{Im}\, G_1^{(0)}$, on ℓ_+ coincide with the corresponding analytical expressions (see [32, 31]), and the values of the normalized drag $F_z/(6\pi\mu v_z)$ are 1.660188 (spindle) and 1.341761 (biconvex lens), which are accurate to within 10^{-6}, compared to the corresponding values obtained by the stream function approach in [32, 31].

Examples 8 and 10 and can be readily extended to the case of two unequal-size spheroids and two fused unequal-size spheres, respectively.

[19] For $\eta \in (0, \pi/2)$, the spindle resembles an "apple," while the biconvex lens is two fused equal spheres; for $\eta = \pi/2$, surfaces of both the spindle and the lens form a sphere; and for $\eta \in (\pi/2, \pi)$, the spindle resembles a "lemon."



(a) $\wp/(2v_z\mu)$, spindle ($\eta = \pi/3$) (b) $\wp/(2v_z\mu)$, biconvex lens ($\eta = \pi/3$)

FIG. 8. *Profiles of the pressure on the surface of the solid spindle and biconvex lens in the axially symmetric translation along the z-axis:* (a) *$\wp/(2v_z\mu)$, spindle with $\eta = \pi/3$ ("apple");* (b) *$\wp/(2v_z\mu)$, biconvex lens with $\eta = \pi/3$ (two fused equal spheres).*

Appendix. Derivation of Cauchy's integral formula for k-harmonically analytic functions.

In this section, we prove Cauchy's integral formula (24) for k-harmonically analytic functions using the approach of Alexandrov and Soloviev [1], who obtained Cauchy's integral formula for 0-harmonically analytic functions.

Let $\mathcal{D}$ be a region symmetric with respect to the z-axis in the rz-plane and having a smooth, positively oriented boundary $\ell = \ell_+ \bigcup \ell_-$, where ℓ_+ is the part of ℓ in the right half of the rz-plane ($r \geq 0$) and ℓ_- is the reflection of ℓ_+ with respect to the z-axis (ℓ_+ is either a closed curve or an open curve with the endpoints lying on the z-axis).

Let $G^{(k)}(\zeta)$ be a k-harmonically analytic function in the region $\mathcal{D}$. In order for the function $G^{(k)}(\zeta)$ to be represented in the form of Cauchy's integral formula,

$$G^{(k)}(\zeta) = \frac{1}{2\pi i}\oint_\ell G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\,d\tau, \qquad \zeta \in \operatorname{int}\mathcal{D}, \tag{84}$$

the following conditions should hold:

(C1) $\mathcal{W}^{(k)}(\zeta,\tau)$ is a k-harmonically analytic function with respect to the variable $\zeta = r + i\,z$; i.e., it solves the Carleman system (3) with respect to ζ.

(C2) The integral (84) is independent of the form of the curve ℓ enclosing the point ζ.

(C3) $\lim_{\tau\to\zeta}[(\tau-\zeta)(\mathcal{W}^{(k)}(\zeta,\tau) - \frac{1}{\tau-\zeta})] = 0$.

With the condition (7) and the symmetry of the contour ℓ with respect to the z-axis, the representation (84) can be reduced to the integral along the curve ℓ_+:

$$G^{(k)}(\zeta) = \frac{1}{2\pi i}\int_{\ell_+} G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\,d\tau + (-1)^k\overline{G^{(k)}(\tau)}\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)d\overline{\tau}. \tag{85}$$

From (85) and (7), it follows that

$$\mathcal{W}^{(k)}\left(-\overline{\zeta},-\overline{\tau}\right)=-\overline{\mathcal{W}^{(k)}(\zeta,\tau)}. \tag{86}$$

We begin with examining condition (C1). Substituting (85) into (6) and using (86), we obtain

$$\frac{1}{2\pi i}\int_{\ell_+}G^{(k)}(\tau)\mathcal{S}(\zeta,\tau)\,d\tau+(-1)^k\overline{G^{(k)}(\tau)}\mathcal{S}\left(\zeta,-\overline{\tau}\right)d\overline{\tau}=0, \tag{87}$$

where

$$\mathcal{S}(\zeta,\tau)=\frac{\partial}{\partial\overline{\zeta}}\mathcal{W}^{(k)}(\zeta,\tau)-\frac{1}{4r}\left((-1)^k\left(2k+1\right)\mathcal{W}^{(k)}\left(-\overline{\zeta},\tau\right)-\mathcal{W}^{(k)}(\zeta,\tau)\right).$$

Consequently, (87) holds if $\mathcal{S}(\zeta,\tau)=0$. This means that $\mathcal{W}^{(k)}(\zeta,\tau)$ satisfies the equation

$$\frac{\partial}{\partial\overline{\zeta}}\mathcal{W}^{(k)}(\zeta,\tau)-\frac{1}{4r}\left((-1)^k\left(2k+1\right)\mathcal{W}^{(k)}\left(-\overline{\zeta},\tau\right)-\mathcal{W}^{(k)}(\zeta,\tau)\right)=0. \tag{88}$$

Now we consider condition (C2). Let

$$\begin{aligned}U_1(\zeta,\tau)&=\operatorname{Re}\left[G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\right], & U_2(\zeta,\tau)&=\operatorname{Re}\left[(-1)^k\overline{G^{(k)}(\tau)}\mathcal{W}^{(k)}(\zeta,-\overline{\tau})\right],\\ V_1(\zeta,\tau)&=\operatorname{Im}\left[G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\right], & V_2(\zeta,\tau)&=\operatorname{Im}\left[(-1)^k\overline{G^{(k)}(\tau)}\mathcal{W}^{(k)}(\zeta,-\overline{\tau})\right];\end{aligned}$$

then the integral (85) can be rewritten as

$$\begin{aligned}G^{(k)}(\zeta)=\frac{1}{2\pi i}\int_{\ell_+}&\left(U_1(\zeta,\tau)+U_2(\zeta,\tau)\right)dr_1+\left(V_2(\zeta,\tau)-V_1(\zeta,\tau)\right)dz_1\\ &+i\left[\left(U_1(\zeta,\tau)-U_2(\zeta,\tau)\right)dz_1+\left(V_2(\zeta,\tau)+V_1(\zeta,\tau)\right)dr_1\right],\end{aligned} \tag{89}$$

where $r_1=\operatorname{Re}\tau$ and $z_1=\operatorname{Im}\tau$. By Green's theorem, (89) is independent of the form of ℓ_+, i.e., it satisfies (C2) if $\frac{\partial}{\partial r_1}(U_1-U_2)-\frac{\partial}{\partial z_1}(V_1+V_2)=0$ and $\frac{\partial}{\partial z_1}(U_1+U_2)-\frac{\partial}{\partial r_1}(V_2-V_1)=0$, which can be restated as

$$\frac{\partial}{\partial\overline{\tau}}\left(G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\right)-\frac{\partial}{\partial\tau}\left((-1)^k\overline{G^{(k)}(\tau)}\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)\right)=0. \tag{90}$$

Figure 9 shows that Green's theorem is applied to the integral (89) along a closed curve consisting of ℓ_+, ℓ'_+, and auxiliary segments. In the case when ℓ_+ is an open curve with the endpoints lying on the z-axis, $\mathcal{W}^{(k)}(\zeta,\tau)$ should vanish at $\operatorname{Re}\tau=0$; see Figure 9(a).

Since the function $G^{(k)}$ satisfies (6), equation (90) reduces to

$$G^{(k)}(\tau)\,\mathcal{T}_1(\zeta,\tau)-(-1)^k\overline{G^{(k)}(\tau)}\,\mathcal{T}_2(\zeta,\tau)=0, \tag{91}$$

where

$$\begin{aligned}\mathcal{T}_1(\zeta,\tau)&=\frac{\partial}{\partial\overline{\tau}}\mathcal{W}^{(k)}(\zeta,\tau)-\frac{1}{4r_1}\left((-1)^k\left(2k+1\right)\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)+\mathcal{W}^{(k)}(\zeta,\tau)\right),\\ \mathcal{T}_2(\zeta,\tau)&=\frac{\partial}{\partial\tau}\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)-\frac{1}{4r_1}\left((-1)^k\left(2k+\right)\mathcal{W}^{(k)}(\zeta,\tau)+\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)\right).\end{aligned} \tag{92}$$

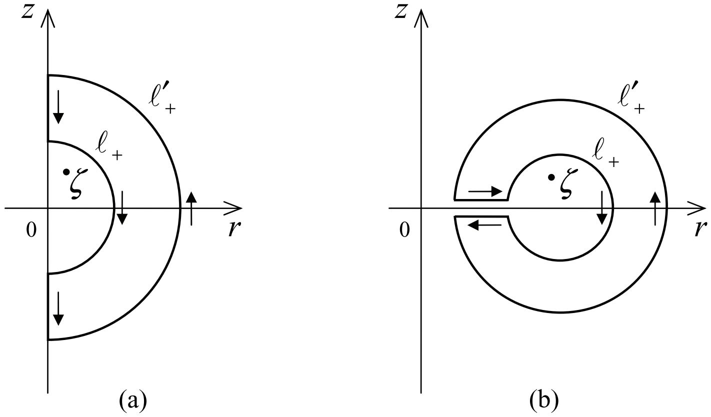


FIG. 9. *Green's theorem is applied to the integral* (89) *along a closed curve consisting of* ℓ_+, ℓ'_+, *and auxiliary segments:* (a) ℓ_+ *is an open curve with the endpoints lying on the* z*-axis;* (b) ℓ_+ *is a closed curve.*

It follows from (86) and (92) that $\mathcal{T}_2(\zeta,\tau) = -\overline{\mathcal{T}_1\left(-\overline{\zeta},\tau\right)}$, and consequently, (91) will hold for any $G^{(k)}$ if $\mathcal{T}_1(\zeta,\tau)=0$; i.e., $\mathcal{W}^{(k)}(\zeta,\tau)$ solves the equation

$$\frac{\partial}{\partial\overline{\tau}}\mathcal{W}^{(k)}(\zeta,\tau) - \frac{1}{4r_1}\Big((-1)^k\,(2k+1)\,\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right) + \mathcal{W}^{(k)}(\zeta,\tau)\Big) = 0. \tag{93}$$

Consequently, $\mathcal{W}^{(k)}(\zeta,\tau)$ solves (88) and (93). Noticing the similarity between (14) and (88), we can represent $\mathcal{W}^{(k)}(\zeta,\tau)$ in the form similar to (9)

$$\mathcal{W}^{(k)}(\zeta,\tau) = \frac{1}{r^k\,|r|}\int_{-\overline{\zeta}}^{\zeta} f(t,\tau)\,(\zeta-t)^{k-\frac{1}{2}}\left(\overline{\zeta}+t\right)^{k+\frac{1}{2}}dt, \tag{94}$$

where $f(t,\tau)$ is an analytic function in the region $\mathcal{D}$ with respect to t except for the points $t=\tau$ and $t=-\overline{\tau}$ lying on the boundary of $\mathcal{D}$. If $\mathcal{D}$ is unbounded, then we also require $f(t,\tau)\sim\mathcal{O}\left(|t|^{-2k-1-\epsilon}\right)$, $\epsilon>0$, at $|t|\to\infty$. However, in contrast to the representation (9), $f(t,\tau)$ is not required to satisfy $f\left(-\overline{\tau}\right)=\overline{f(\tau)}$. Making two branch cuts connecting each of the points ζ and $-\overline{\zeta}$ with the boundary $\mathcal{D}$, we can show that (94) is independent of the form of a simple curve that connects $-\overline{\zeta}$ and ζ in $\mathcal{D}$ and contains neither τ nor $-\overline{\tau}$ (similar to that in Proposition 1).

Substituting (94) into (93), we obtain an equation for $f(t,\tau)$:

$$\frac{\partial}{\partial\overline{\tau}}f(t,\tau) - \frac{1}{4r_1}\left((-1)^k\,(2k+1)\,f\left(t,-\overline{\tau}\right) + f(t,\tau)\right) = 0. \tag{95}$$

The similarity between (14) and (93) implies that the simplest solution to (95) can be determined by

$$f(t,\tau) = \frac{\left(k+\frac{1}{2}\right)r_1^k\,|r_1|}{(\tau-t)^{k+\frac{3}{2}}\left(\overline{\tau}+t\right)^{k+\frac{1}{2}}},$$

which is an analytic function with respect to t in the whole plane except for the points $t=\tau$ and $t=-\overline{\tau}$. Thus, $\mathcal{W}^{(k)}(\zeta,\tau)$, represented by (94), takes the form

$$\mathcal{W}^{(k)}(\zeta,\tau) = \left(k+\frac{1}{2}\right)\left(\frac{r_1}{r}\right)^k\left|\frac{r_1}{r}\right|\int_{-\overline{\zeta}}^{\zeta}\frac{1}{(\tau-t)(\zeta-t)}\left[\frac{\left(\overline{\zeta}+t\right)(\zeta-t)}{\left(\overline{\tau}+t\right)(\tau-t)}\right]^{k+\frac{1}{2}}dt. \tag{96}$$

It is seen that (96) satisfies (86) and the condition $\mathcal{W}^{(k)}(\zeta,\tau)=0$ at $\operatorname{Re}\tau=r_1=0$. First, we evaluate (96) and then show that (C3) holds. In order to evaluate (96), we assume that $r>0$ and consider two cases: $r_1>0$ and $r_1<0$, which correspond to $\mathcal{W}^{(k)}(\zeta,\tau)$ and $\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)$ in (85), respectively.

With four branch cuts connecting each of the points ζ, $-\overline{\zeta}$, τ, and $-\overline{\tau}$ with infinite points, the integral (96) does not depend on a curve connecting $-\overline{\zeta}$ and ζ, and consequently, for $\tau\neq\zeta$, we can choose a curve $t=t(x)$ determined by

$$t(x)=\frac{\zeta\left(\overline{\tau}-\overline{\zeta}\right)-\overline{\zeta}\,\overline{\tau}\cos^2 x}{\overline{\tau}+\zeta-\left(\zeta+\overline{\zeta}\right)\sin^2 x},\qquad x\in[0,\pi/2],\tag{97}$$

for which

$$\frac{\left(t(x)+\overline{\zeta}\right)\left(\zeta+\overline{\tau}\right)}{\left(t(x)+\overline{\tau}\right)\left(\zeta+\overline{\zeta}\right)}=\sin^2 x,\qquad x\in[0,\pi/2].$$

In the case when $r_1>0$, the integral (96) along the curve (97) reduces to

$$\mathcal{W}^{(k)}(\zeta,\tau)=2^{-2k}(2k+1)\frac{\lambda(\zeta,\tau)^{2k}\left(1-\lambda(\zeta,\tau)^2\right)}{\tau-\zeta}\left|\frac{\tau+\overline{\tau}}{\zeta+\overline{\tau}}\right|\int_0^{\frac{\pi}{2}}\frac{(\sin[2x])^{2k}\sin^2 x}{\left(1-\lambda^2\sin^2 x\right)^{k+\frac{3}{2}}}\,dx,\tag{98}$$

where $\lambda(\zeta,\tau)\in[0,1]$ is a real-valued function defined by (27). With the definition of the hypergeometric function (26), the integral (98) takes the form

$$\mathcal{W}^{(k)}(\zeta,\tau)=\frac{\Omega_+^{(k)}(\zeta,\tau)}{\tau-\zeta},\qquad \operatorname{Re}\zeta\geq 0,\quad \operatorname{Re}\tau\geq 0,$$

where $\Omega_+^{(k)}(\zeta,\tau)$ is determined by (25a).

In the case when $r_1<0$, we have

$$\begin{aligned}&\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)\\&=(-1)^{k+1}\left(k+\frac{1}{2}\right)\left(\frac{r_1}{r}\right)^k\left|\frac{r_1}{r}\right|\int_{-\overline{\zeta}}^{\zeta}\frac{1}{\left(\overline{\tau}+t\right)\left(\zeta-t\right)}\left[\frac{\left(\overline{\zeta}+t\right)\left(\zeta-t\right)}{\left(\overline{\tau}+t\right)\left(\tau-t\right)}\right]^{k+\frac{1}{2}}dt.\end{aligned}\tag{99}$$

Using the same curve (97), we obtain

$$\mathcal{W}^{(k)}\left(\zeta,-\overline{\tau}\right)=(-1)^{k+1}\frac{\Omega_-^{(k)}(\zeta,\tau)}{\overline{\tau}+\zeta},\qquad \operatorname{Re}\zeta\geq 0,\quad \operatorname{Re}\tau\geq 0,$$

where $\Omega_-^{(k)}(\zeta,\tau)$ is determined by (25b).

Now we are ready to verify (C3). Note that $\lambda\to 1-$ as $\tau\to\zeta$. Using the asymptotic representation for the hypergeometric functions

$$\begin{aligned}\mathbb{F}\left(k+\tfrac{3}{2},k+\tfrac{3}{2},2(k+1),\lambda^2\right)=&\frac{\Gamma(2(k+1))}{\left[\Gamma\left(k+\frac{3}{2}\right)\right]^2}\,\frac{1}{1-\lambda^2}+\frac{\Gamma(2(k+1))}{\left[\Gamma\left(k+\frac{1}{2}\right)\right]^2}\ln\left(1-\lambda^2\right)\\&+\mathcal{O}\left(1\right)\ \text{ as }\lambda\to 1-,\end{aligned}$$

$$\begin{aligned}\mathbb{F}\left(k+\tfrac{1}{2},k+\tfrac{3}{2},2(k+1),\lambda^2\right)=&-\frac{\Gamma(2(k+1))}{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(k+\frac{3}{2}\right)}\\&\times\left(\ln(1-\lambda^2)+2\left(\frac{1}{2k+1}+\gamma+\psi\left(k+\tfrac{1}{2}\right)\right)\right)\\&+\mathcal{O}\left((1-\lambda^2)\right)\ \text{ as }\lambda\to 1-,\end{aligned}$$

where γ is the Euler constant and $\psi(\cdot)$ is the digamma function, we have $\lim_{\tau\to\zeta}\Omega_+^{(k)}(\zeta,\tau) = 1$. Similarly, the asymptotic form of $\Omega_-^{(k)}(\zeta,\tau)$ as $\tau\to\zeta$ is determined by (28). In other words, the function $\Omega_-^{(k)}(\zeta,\tau)$ has a logarithmic singularity at $\tau=\zeta$,[†] and consequently, the difference $\mathcal{W}^{(k)}(\zeta,\tau)-\frac{1}{\tau-\zeta}$ has only integrable (logarithmic) singularity.

Finally, we show that Cauchy's integral formula (84) holds. Let C_ρ be the positively oriented circle of the radius ρ with the center at ζ; then the integral (85) along ℓ_+ equals the integral over C_ρ:

$$\begin{aligned}&\frac{1}{2\pi i}\int_{\ell_+} G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\,d\tau+(-1)^k\overline{G^{(k)}(\tau)}\mathcal{W}^{(k)}\,(\zeta,-\overline{\tau})\,d\overline{\tau}\\ &=\frac{1}{2\pi i}\oint_{C_\rho}\underbrace{G^{(k)}(\tau)\,\frac{\Omega_+^{(k)}(\zeta,\tau)}{\tau-\zeta}\,d\tau}_{I_1}-\underbrace{\overline{G^{(k)}(\tau)}\,\frac{\Omega_-^{(k)}(\zeta,\tau)}{\overline{\tau}+\zeta}\,d\overline{\tau}}_{I_2},\quad \zeta\in\operatorname{int}\mathcal{D}.\end{aligned}\tag{100}$$

Indeed, we can form a closed simple curve, consisting of ℓ_+, the circle C_ρ, and auxiliary segments (see Figure 10), that bounds the region in which $G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)$ has continuous partial derivatives, and then apply Green's theorem. Let $\tau=\zeta+\rho\,\mathrm{e}^{i\varphi}$. Since $\Omega_-^{(k)}(\zeta,\tau)$ has only logarithmic singularity as $\tau\to\zeta$, the integral I_2 in (100) vanishes at $\rho\to 0$; and since $\lim_{\tau\to\zeta}\Omega_+^{(k)}(\zeta,\tau)=1$, the integral I_1 reduces to $G^{(k)}(\zeta)$ when $\rho\to 0$, which completes the proof of Cauchy's integral formula for k-harmonically analytic functions.

Acknowledgment. We are grateful to the anonymous referees for their valuable comments and suggestions, which helped to improve the quality of the paper.

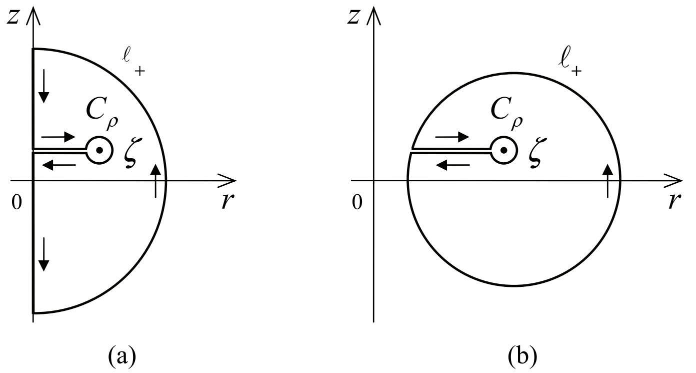


FIG. 10. *A closed simple curve, consisting of ℓ_+, the circle C_ρ, and auxiliary segments, bounds the region in which $G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)$ has continuous partial derivatives:* (a) *ℓ_+ is an open curve with the endpoints lying on the z-axis;* (b) *ℓ_+ is a closed curve.*

REFERENCES

[1] A. Ya. Alexandrov and Yu. I. Soloviev, *Three-Dimensional Problems of the Theory of Elasticity*, Nauka, Moscow, 1978 (in Russian).

[2] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 2000.

[†] Note that there is no singularity at $\tau=\zeta$ if $\operatorname{Re}\zeta=0$.

[3] H. Bateman and A. Erdelyi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[4] L. Bers, *The expansion theorem for sigma-monogenic functions*, Amer. J. Math., 72 (1950), pp. 705–712.

[5] L. Bers, *Theory of Pseudo-Analytic Functions*, Institute for Mathematics and Mechanics, New York University, New York, 1953.

[6] L. Bers, *An outline of the theory of pseudoanalytic functions*, Bull. Amer. Math. Soc., 62 (1956), pp. 291–331.

[7] L. Bers and A. Gelbart, *On a class of differential equations in mechanics of continua*, Quart. Appl. Math., 1 (1943), pp. 168–188.

[8] L. Bers and A. Gelbart, *On a class of functions defined by partial differential equations*, Trans. Amer. Math. Soc., 56 (1944), pp. 67–93.

[9] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, New York, 1966.

[10] S. L. Goren and M. E. O'Neill, *Asymmetric creeping motion of an open torus*, J. Fluid Mech., 101 (1980), pp. 97–110.

[11] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Springer-Verlag, New York, 1983.

[12] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea Publishing, New York, 1955.

[13] V. V. Kravchenko, *On the relationship between p-analytic functions and Schrödinger equation*, Z. Anal. Anwendungen, 24 (2005), pp. 487–496.

[14] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1945.

[15] N. N. Lebedev, *The functions associated with a ring of oval cross section*, Technical Physics USSR, 1 (1937), pp. 3–24.

[16] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953.

[17] N. I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*, 1st ed., Springer-Verlag, New York, 1977.

[18] N. I. Muskhelishvili, *Singular Integral Equations: Boundary Problems of Function Theory and Their Applications to Mathematical Physics*, 2nd ed., Dover, New York, 1992.

[19] W. H. Pell and L. E. Payne, *On Stokes flow about a torus*, Mathematika, 7 (1960), pp. 78–92.

[20] G. N. Položii, *Theory and Application of p-Analytic and (p, q)-Analytic Functions*, 2nd ed., Naukova Dumka, Kiev, 1973 (in Russian).

[21] C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge University Press, New York, 1992.

[22] C. Pozrikidis, *Introduction to Theoretical and Computational Fluid Dynamics*, Oxford University Press, New York, 1996.

[23] M. Stimson and G. B. Jeffery, *The motion of two-spheres in a viscous fluids*, Proc. Roy. Soc. London Ser. A, 111 (1926), pp. 110–116.

[24] I. N. Vekua, *Generalized Analytic Functions*, Pergamon Press, Oxford, UK, 1962.

[25] S. Wakiya, *Mutual interaction of two spheroids sedimenting in a viscous fluid*, J. Phys. Soc. Japan, 20 (1965), pp. 1502–1514.

[26] S. Wakiya, *Slow motion of a viscous fluid around two spheres*, J. Phys. Soc. Japan, 22 (1967), pp. 1101–1109.

[27] S. Wakiya, *On the exact solution of the Stokes equations for a torus*, J. Phys. Soc. Japan, 37 (1974), pp. 780–783.

[28] A. Weinstein, *Generalized axially symmetric potential theory*, Bull. Amer. Math. Soc., 59 (1953), pp. 20–38.

[29] M. Zabarankin, *The framework of k-harmonically analytic functions for three-dimensional Stokes flow problems, Part* II, SIAM J. Appl. Math., 69 (2008), pp. 881–907.

[30] M. Zabarankin and P. Krokhmal, *Generalized analytic functions in* 3*D Stokes flows*, Quart. J. Mech. Appl. Math., 60 (2007), pp. 99–123.

[31] M. Zabarankin and A. F. Ulitko, *Hilbert formulae for r-analytic functions and Stokes flow about a biconvex lens*, Quart. Appl. Math., 64 (2006), pp. 663–693.

[32] M. Zabarankin and A. F. Ulitko, *Hilbert formulae for r-analytic functions in the domain exterior to spindle*, SIAM J. Appl. Math., 66 (2006), pp. 1270–1300.

SIAM J. Appl. Math.
Vol. 69, No. 3, pp. 881–907

THE FRAMEWORK OF k-HARMONICALLY ANALYTIC FUNCTIONS FOR THREE-DIMENSIONAL STOKES FLOW PROBLEMS, PART II*

MICHAEL ZABARANKIN†

Abstract. A solution form representing the velocity field and pressure for asymmetric three-dimensional (3D) Stokes flows has been constructed in terms of three k-harmonically analytic functions. It has also been shown that it uniquely determines an external velocity field vanishing at infinity. With the obtained solution form, problems of 3D Stokes flows due to asymmetric motions of solid bodies of revolution have been reduced to boundary-value problems for the three k-harmonically analytic functions, and the resisting force and torque, exerted on bodies in corresponding motions, have been expressed in terms of the k-harmonically analytic functions entering the solution form. For regions, in which Laplace's equation admits separation of variables, the boundary-value problems can be solved in closed form via series or integral representations of k-harmonically analytic functions in corresponding curvilinear coordinates. This approach has been demonstrated for asymmetric translation and rotation of solid sphere and solid prolate and oblate spheroids. As the second approach, the boundary-value problems have been reduced to integral equations based on Cauchy's integral formula for k-harmonically analytic functions. As an illustration, the integral equations have been solved for asymmetric translation and rotation of solid bispheroids and a solid torus of elliptical cross-section for various values of a geometrical parameter.

Key words. asymmetric Stokes flows, generalized analytic functions, exact solution, generalized Cauchy's integral formula, integral equation

AMS subject classifications. 30E20, 35Q15, 35Q30, 76D07

DOI. 10.1137/080715925

1. Introduction.

This article presents the second part of the developed two-part framework of k-harmonically analytic functions in application to three-dimensional (3D) Stokes flows. In the first part [32], we obtained Cauchy's integral formula for k-harmonically analytic functions and constructed a solution form for the velocity field and pressure for axially symmetric Stokes flows in terms of two 0-harmonically analytic functions. This work extends all the results obtained in [32] to *asymmetric* 3D Stokes flows.

1.1. Stokes equations.

The behavior of steady flows of a viscous incompressible fluid under the assumption of zero (low) Reynolds number (so-called *Stokes creeping flows*) is described by the Stokes equations

$$\mu\,\boldsymbol{\Delta}\mathbf{u} = \operatorname{grad}\wp, \qquad \operatorname{div}\mathbf{u} = 0, \tag{1}$$

where $\mathbf{u}$ is the fluid velocity field, $\wp$ is the pressure in the fluid, μ is the shear viscosity, and $\boldsymbol{\Delta}\mathbf{u} \equiv \operatorname{grad}(\operatorname{div}\mathbf{u}) - \operatorname{curl}(\operatorname{curl}\mathbf{u})$. In (1), the first equation is known as the Stokes creeping flow equation and the second one is the equation of continuity; see, e.g., [10, 9]. The model (1) can also be represented in the form $\operatorname{grad}\wp = -\mu\operatorname{curl}(\operatorname{curl}\mathbf{u})$, $\operatorname{div}\mathbf{u} = 0$, whence it follows that the pressure $\wp$ and vorticity $\boldsymbol{\omega} = \operatorname{curl}\mathbf{u}$ are related

*Received by the editors February 18, 2008; accepted for publication (in revised form) August 12, 2008; published electronically December 31, 2008. This research was partially supported by AFOSR grant FA9550-07-1-0401.

http://www.siam.org/journals/siap/69-3/71592.html

†Department of Mathematical Sciences, Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030 (mzabaran@stevens.edu).

scalar and vectorial potentials, i.e., they satisfy $\operatorname{grad}\wp = -\mu \operatorname{curl}\boldsymbol{\omega}$ with $\operatorname{div}\boldsymbol{\omega} = 0$; see [33, 32].

In this work, we consider asymmetric 3D Stokes flows due to motion of a solid body of revolution. By asymmetric motion of the body, we will understand translation along and rotation around axes transversal to the axis of revolution. Let S be the surface of the body and let the body move in the fluid with the velocity $\mathbf{u}_0$. The no-slip boundary conditions on the surface of the body are formulated by

$$\mathbf{u} = \mathbf{u}_0 \quad \text{on} \quad S, \tag{2}$$

and also the velocity field $\mathbf{u}$ and pressure $\wp$ vanish at infinity:

$$\mathbf{u}|_{\infty} = 0, \qquad \wp|_{\infty} = 0. \tag{3}$$

It is known that the Stokes flow problem (1)–(3) has a unique solution; see, e.g., [2, section 4.9]. This fact follows from the following proposition.

Proposition 1 (homogeneous Stokes flow problem). *The problem* (1) *and* (3) *with zero boundary conditions* $\mathbf{u}|_S = 0$ *has only a zero solution, i.e.,* $\mathbf{u} \equiv 0$*, in the corresponding outer region.*

Proof. The proof can be found in [2, section 4.9]. First, it is shown that the Stokes equations with (3) and $\mathbf{u}|_S = 0$ imply $\operatorname{curl}\mathbf{u} = 0$, and then it is proved that the problem $\operatorname{div}\mathbf{u} = 0$, $\operatorname{curl}\mathbf{u} = 0$ subject to (3) and $\mathbf{u}|_S = 0$ has only a zero solution. For the last problem, special attention is paid to multiply connected regions. For details, see [2, sections 2.8, 4.9]. □

Proposition 1 will be central in establishing uniqueness of solutions for the velocity field represented in terms of k-harmonically analytic functions.

Let (r, φ, z) be the cylindrical coordinate system with the basis $(\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{k})$. Without loss of generality, we can represent the velocity field $\mathbf{u}$ and pressure $\wp$ by

$$\begin{aligned}\mathbf{u}(r,\varphi,z) = \sum_{k=0}^{\infty} u_r^{(k)}(r,z) \left\{ \begin{array}{c} \cos \\ \sin \end{array} (k\varphi) \right\} \mathbf{e}_r + u_\varphi^{(k)}(r,z) \left\{ \begin{array}{c} \sin \\ -\cos \end{array} (k\varphi) \right\} \mathbf{e}_\varphi \\ + u_z^{(k)}(r,z) \left\{ \begin{array}{c} \cos \\ \sin \end{array} (k\varphi) \right\} \mathbf{k}\end{aligned} \tag{4}$$

and

$$\wp(r,\varphi,z) = \sum_{k=0}^{\infty} \wp^{(k)}(r,z) \left\{ \begin{array}{c} \cos \\ \sin \end{array} (k\varphi) \right\}, \tag{5}$$

where the choice of either upper or lower functions in the curly brackets depends on whether $\mathbf{u}$ and $\wp$ are even or odd functions with respect to φ. If $\mathbf{u}$ and $\wp$ are neither odd nor even, they are represented by sums of series, corresponding to even and odd parts of $\mathbf{u}$ and $\wp$, respectively.

For convenience, we denote

$$u_1^{(k)} = u_r^{(k)} - u_\varphi^{(k)}, \qquad u_2^{(k)} = u_r^{(k)} + u_\varphi^{(k)}. \tag{6}$$

The first equation in (1) reduces to a series of equations for $k \in \mathbb{Z}_0^+$:

$$\mu\,\Delta_{k-1} u_1^{(k)} = \left(\frac{\partial}{\partial r} + \frac{k}{r}\right)\wp^{(k)}, \quad \mu\,\Delta_{k+1} u_2^{(k)} = \left(\frac{\partial}{\partial r} - \frac{k}{r}\right)\wp^{(k)}, \quad \mu\,\Delta_k u_z^{(k)} = \frac{\partial}{\partial z}\wp^{(k)}, \tag{7}$$

where Δ_k denotes the so-called k-harmonic operator:

$$\Delta_k \equiv \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2} - \frac{k^2}{r^2}. \tag{8}$$

Similarly, the continuity equation $\operatorname{div} \mathbf{u} = 0$ reduces to

$$\left(\frac{\partial}{\partial r} - \frac{k-1}{r}\right) u_1^{(k)} + \left(\frac{\partial}{\partial r} + \frac{k+1}{r}\right) u_2^{(k)} + 2\frac{\partial}{\partial z} u_z^{(k)} = 0. \tag{9}$$

Since $\wp(r,\varphi,z)$ satisfies $\operatorname{div}(\operatorname{grad}\wp) = 0$, the function $\wp^{(k)}(r,z)$ is k-harmonic, i.e.,

$$\Delta_k\, \wp^{(k)} = 0, \tag{10}$$

and consequently, it follows from (7) and (10) that

$$\Delta_{k-1}^2\, u_1^{(k)} = 0, \qquad \Delta_{k+1}^2\, u_2^{(k)} = 0, \qquad \Delta_k^2\, u_z^{(k)} = 0. \tag{11}$$

In the axially symmetric case, all equations (7)–(11) are considered for $k = 0$, and in the case of asymmetric motion of a body of revolution, it is sufficient to consider (7)–(11) for only $k = 1$.

There are two well-known solution forms for the Stokes flow problem (1)–(3): one in terms of a stream function, which identically solves (9) for $k = 0$ and is applicable only for the axially symmetric case (see [9]), and the other due to Dean and O'Neill [6], which should satisfy the continuity equation. Both solution forms are mainly used for canonical regions, i.e., those in which Laplace's equation admits separation of variables. Stream-function solutions to the axially symmetric Stokes flow problem were obtained for particles of virtually all known canonical shapes, including *sphere* [23], *prolate* and *oblate spheroids* [14, 9], *circular disk* [14, 9], *spherical cap* [16, 5, 25], *two spheres* [22], *torus* [17, 28], *spindle* [18, 35], and *lens* [16, 34]. As for asymmetric problems, exact solutions in the form of Dean and O'Neill's were also constructed for the majority of particles of canonical shape as follows: *two spheres* [15, 27, 13], *torus* [28, 8, 24], *two fused equal spheres* [30], and *spindle* [31].

For particles of arbitrary shape, exact solutions to 3D Stokes flow problems can be obtained by integral equation approaches [20, 21]. A well-known approach uses integral representations for harmonic functions via Green's functions [21, 29]. Yet another approach can be contemplated based on the fact that for an incompressible isotropic elastic medium, the Lamé equation formally corresponds to the Stokes model (1) (see [25, 35]), and that in the axially symmetric case, the displacement vector $\mathbf{u}$ can be represented by a generalized Kolosov–Muskhelishvili formula in terms of two p-analytic functions (with $p = r$) and their derivatives [19, 1]. However, to the best of our knowledge, no representation of an *asymmetric* 3D velocity field in terms of either generalized analytic or pseudoanalytic functions has been encountered in the existing literature. Coupled with a corresponding generalized Cauchy's integral formula, such a representation would allow one to reduce 3D Stokes flow problems to integral equations.

As a central result of this work, we obtain a solution form for the velocity field and pressure for asymmetric 3D Stokes flow problems in terms of three k-harmonically analytic functions. This solution form is similar to Goursat's formula representing a solution to a 2D biharmonic equation via two ordinary analytic functions. Its main advantage is that, in contrast to Dean and O'Neill's solution form, it identically satisfies the continuity equation, and in contrast to generalized Kolosov–Muskhelishvili formulae, it does not involve derivatives of k-harmonically analytic functions.[1] It is applicable to obtaining closed-form analytical solutions for the canonical regions as well as to obtaining exact solutions via integral equations for arbitrary regions.

1.2. k-harmonically analytic functions. Here we define only k-harmonically analytic functions and refer the reader to the first part of this work [32], in which we derived a Cauchy's integral formula for k-harmonically analytic functions and constructed series representations for k-harmonically analytic functions for the regions exterior to a sphere and prolate and oblate spheroids.

For each $k \in \mathbb{Z}_0^+$, k-harmonically analytic functions $G^{(k)}(r,z) = U^{(k)}(r,z) + i\,V^{(k+1)}(r,z)$, $i = \sqrt{-1}$, constitute a class of generalized analytic functions [3, 4, 19, 26] and are determined by a particular case of the Bers–Vekua system

$$\left(\frac{\partial}{\partial r} - \frac{k}{r}\right) U^{(k)} = \frac{\partial}{\partial z} V^{(k+1)}, \qquad \frac{\partial}{\partial z}\, U^{(k)} = -\left(\frac{\partial}{\partial r} + \frac{k+1}{r}\right) V^{(k+1)}. \tag{12}$$

In [33], we showed that (12) arises from an asymmetric 3D case of the relationship

$$\operatorname{grad}\phi = -\operatorname{curl}\boldsymbol{\Lambda}, \quad \operatorname{div}\boldsymbol{\Lambda} = 0 \tag{13}$$

for a scalar field ϕ and vectorial field $\boldsymbol{\Lambda}$, which in view of (13) are called related potentials. In the 2D case in Cartesian coordinates, (13) reduces to the classical Cauchy–Riemann system for ordinary analytic functions, and in the axially symmetric 3D case in the cylindrical coordinates (r, φ, z),[2] (13) defines so-called r-analytic functions; see [33]. In the asymmetric 3D case, (13) reduces to (12), which relates kth harmonics of ϕ and $\boldsymbol{\Lambda}$ in the cylindrical coordinates with respect to the angular coordinate φ.

It follows from (12) that $U^{(k)}$ and $V^{(k+1)}$ are k-harmonic and $(k+1)$-harmonic functions, respectively:

$$\Delta_k U^{(k)} = 0 \quad \text{and} \quad \Delta_{k+1} V^{(k+1)} = 0.$$

To emphasize this fact, we call $G^{(k)}$ satisfying (12) a *k-harmonically analytic* function. In particular, for $k = 0$, the system (12) defines the class of r-analytic functions; see [33, 34, 35].

Let $\zeta = r + i\,z$ be a complex variable. Introducing the derivative $\frac{\partial}{\partial\bar\zeta} = \frac{1}{2}\left(\frac{\partial}{\partial r} + i\,\frac{\partial}{\partial z}\right)$, we can represent the system (12) in the form

$$\frac{\partial G^{(k)}}{\partial\bar\zeta} = \frac{1}{4r}\left((2k+1)\,\overline{G^{(k)}} - G^{(k)}\right). \tag{14}$$

[1] For example, in the asymmetric problem of Stokes flows due to transversal translation of a solid torus or two solid spheres (see [8, 27]), the continuity equation with Dean and O'Neill's solution form reduces to a second-order difference equation. This procedure relies substantially on the peculiar properties of the special functions associated with the bispherical and toroidal coordinates and is unlikely to be extended to the problem with a body of arbitrary shape. Also, the integral equation for the axially symmetric problem of an elastic isotropic medium obtained via generalized Kolosov–Muskhelishvili formulae (see [1, equation (46.40)]) involves a derivative of a generalized Cauchy's kernel, which complicates numerical analysis of the equation.

[2] In this case, the z-axis is the axis of revolution, and ϕ and $\boldsymbol{\Lambda}$ are independent of the angular coordinate φ.

Without assuming analyticity, we will formally write $G^{(k)}(\zeta) = G^{(k)}(r,z)$. A k-harmonically analytic function can be defined for $r < 0$ by introducing the symmetry condition

$$G^{(k)}\left(-\overline{\zeta}\right) = (-1)^k\overline{G^{(k)}(\zeta)}. \tag{15}$$

This condition is dictated by the representation of the velocity field in terms of k-harmonically analytic functions.

The paper is organized into three sections. Section 2 constructs the solution form for asymmetric Stokes flow problems in terms of three k-harmonically analytic functions; proves that the solution form uniquely determines an external velocity field vanishing at infinity; expresses the resisting force and torque, exerted on a solid body of revolution in the Stokes flow in terms of the k-harmonically analytic functions entering the solution form; and reduces Stokes flow problems to integral equations based on Cauchy's integral formula for k-harmonically analytic functions. Section 3 demonstrates the solution form in obtaining analytical solutions to the problems of Stokes flows due to transversal translation and rotation of a solid sphere and solid prolate and oblate spheroids. It also solves the integral equations for transversal translation and rotation of solid bispheroids and a solid torus of elliptical cross-section for various values of a geometrical parameter. The appendix presents two auxiliary results dealing with implementation of the necessary and sufficient condition for a function to be k-harmonically analytic in an outer region and vanishing at infinity.

2. Stokes equations in the asymmetric case.

This section constructs a representation for the velocity field and pressure for asymmetric 3D Stokes flows in terms of three k-harmonically analytic functions.

To simplify notation, we will write a function of the variables of r and z as the function of ζ without assuming analyticity.

PROPOSITION 2 (representation for the velocity field and pressure). *Let $G_1^{(k-1)}(\zeta) = U_1^{(k-1)}(\zeta) + i\,V_1^{(k)}(\zeta)$, $G_2^{(k-1)}(\zeta) = U_2^{(k-1)}(\zeta) + i\,V_2^{(k)}(\zeta)$, and $G_3^{(k)}(\zeta) = U_3^{(k)}(\zeta) + i\,V_3^{(k+1)}(\zeta)$ be k-harmonically analytic functions satisfying* (12) *for $k-1$, $k-1$, and k, respectively, and vanishing at $|\zeta| \to \infty$. Then for $k \geq 1$, the components of the velocity field in the asymmetric case of the Stokes equations can be represented in the form*

(16)

$$2\left(u_z^{(k)}(\zeta) + i\,u_r^{(k)}(\zeta)\right) = \left(\frac{k+1}{2k-1}\,r - i\frac{k-2}{2k-1}\,z\right)G_1^{(k-1)}(\zeta) + i\,G_2^{(k-1)}(\zeta) + G_3^{(k)}(\zeta),$$

$$2u_\varphi^{(k)}(\zeta) = \operatorname{Im}\left[\frac{k-2}{2k-1}(r+i\,z)\,G_1^{(k-1)}(\zeta) - i\,G_2^{(k-1)}(\zeta) + G_3^{(k)}(\zeta)\right],$$

and the kth harmonic of the pressure is determined by

$$\wp^{(k)}(\zeta) = \mu\,\operatorname{Im} G_1^{(k-1)}(\zeta), \quad k \geq 1. \tag{17}$$

Proof. For $k \geq 1$, the components $u_1^{(k)}$, $u_2^{(k)}$, and $u_z^{(k)}$ satisfy (7) and (9). We denote $\wp^{(k)}(\zeta) = V_1^{(k)}(\zeta)$, where $V_1^{(k)}$ is a k-harmonic function. Since $\wp^{(k)}(\zeta)$ vanishes at $|\zeta| \to \infty$, there exists a unique $(k-1)$-harmonic function $U_1^{(k-1)}(\zeta)$ also vanishing at $|\zeta| \to \infty$ such that $U_1^{(k-1)}$ and $V_1^{(k)}$ form a $(k-1)$-harmonically analytic function $G_1^{(k-1)}(\zeta) = U_1^{(k-1)}(\zeta) + i\,V_1^{(k)}(\zeta)$; see [33, Proposition 1]. Using the system (12) for

$k-1$, we can restate the Stokes equations (7) as

$$\begin{aligned}
\Delta_{k-1} u_1^{(k)} &= \left(\frac{\partial}{\partial r} + \frac{k}{r}\right) V_1^{(k)} = -\frac{\partial}{\partial z} U_1^{(k-1)}, \\
\Delta_{k+1} u_2^{(k)} &= \left(\frac{\partial}{\partial r} - \frac{k}{r}\right) V_1^{(k)}, \\
\Delta_k u_z^{(k)} &= \frac{\partial}{\partial z} V_1^{(k)} = \left(\frac{\partial}{\partial r} - \frac{k-1}{r}\right) U_1^{(k-1)}.
\end{aligned} \tag{18}$$

With the identities

$$\begin{aligned}
&\Delta_{k-1}\left(r\, V_1^{(k)}\right) = 2\left(\frac{\partial}{\partial r} + \frac{k}{r}\right) V_1^{(k)}, \qquad \Delta_{k-1}\left(z\, U_1^{(k-1)}\right) = 2\frac{\partial}{\partial z} U_1^{(k-1)}, \\
&\Delta_{k+1}\left(r\, V_1^{(k)}\right) = 2\left(\frac{\partial}{\partial r} - \frac{k}{r}\right) V_1^{(k)}, \qquad \Delta_{k}\left(r\, U_1^{(k-1)}\right) = 2\left(\frac{\partial}{\partial r} - \frac{k-1}{r}\right) U_1^{(k-1)}, \\
&\Delta_k\left(z\, V_1^{(k)}\right) = 2\frac{\partial}{\partial z} V_1^{(k)},
\end{aligned} \tag{19}$$

equations (18) are integrated, and the components $u_1^{(k)}$, $u_2^{(k)}$, and $u_z^{(k)}$ are represented in the form

$$\begin{aligned}
u_1^{(k)}(\zeta) &= a\, r\, V_1^{(k)}(\zeta) + b\, z\, U_1^{(k-1)}(\zeta) + U_2^{(k-1)}(\zeta), \\
u_2^{(k)}(\zeta) &= \frac{1}{2}\, r\, V_1^{(k)}(\zeta) + V_3^{(k+1)}(\zeta), \\
u_z^{(k)}(\zeta) &= c\, z\, V_1^{(k)}(\zeta) + d\, r\, U_1^{(k-1)}(\zeta) + W^{(k)}(\zeta),
\end{aligned} \tag{20}$$

where a, b, c, and d are real-valued constants and $U_2^{(k-1)}$, $V_3^{(k+1)}$, and $W^{(k)}$ are arbitrary functions that satisfy

$$\Delta_{k-1} U_2^{(k-1)} = 0, \qquad \Delta_{k+1} V_3^{(k+1)} = 0, \qquad \Delta_k W^{(k)} = 0$$

and vanish at $|\zeta| \to \infty$. Substituting (20) into (18), we have

$$2a - 2b = 1, \qquad 2c + 2d = 1. \tag{21}$$

Then substituting (20) into (9) and using (12) for $k-1$, we obtain

$$\left(\frac{\partial}{\partial r} - \frac{k-1}{r}\right) U_2^{(k-1)} + \left(\frac{\partial}{\partial r} + \frac{k+1}{r}\right) V_3^{(k+1)} + 2\frac{\partial}{\partial z} W^{(k)} = 0, \tag{22}$$

provided that

$$-a + 2d = \frac{1}{2}, \qquad b + 2c = 0, \qquad 2(1-k)a + 2c = -1. \tag{23}$$

As in the axially symmetric case [32, Proposition 7], (21) and (23) are dependent. Indeed, adding $a - b = 1/2$, $-a + 2d = 1/2$, and $b + 2c = 0$, we have $2c + 2d = 1$. Excluding, for example, $2c + 2d = 1$ from (21) and (23), we obtain the unique solution to the remaining four equations:

$$a = \frac{3}{2(2k-1)}, \quad b = -\frac{k-2}{2k-1}, \quad c = \frac{k-2}{2(2k-1)}, \quad d = \frac{k+1}{2(2k-1)},$$

where $2k - 1 \neq 0$ for any integer k.

Equation (22) becomes an identity for $W^{(k)}(\zeta) = \frac{1}{2}(U_3^{(k)}(\zeta) - V_2^{(k)}(\zeta))$, where $V_2^{(k)}(\zeta)$ is the imaginary part of the $(k-1)$-harmonically analytic function $G_2^{(k-1)}(\zeta) = U_2^{(k-1)}(\zeta) + i\,V_2^{(k)}(\zeta)$ and $U_3^{(k)}(\zeta)$ is the real part of the k-harmonically analytic function $G_3^{(k)}(\zeta) = U_3^{(k)}(\zeta) + i\,V_3^{(k+1)}(\zeta)$. Under the condition that $G_2^{(k-1)}$ and $G_3^{(k)}$ vanish at infinity, $G_2^{(k-1)}$ and $G_3^{(k)}$ are uniquely determined by $U_2^{(k-1)}$ and $V_3^{(k+1)}$, respectively; see [33, Proposition 1].

Consequently, the representation (20) takes the form

(24)
$$\begin{aligned}
u_1^{(k)}(\zeta) &= \frac{3}{2(2k-1)}\,r\,V_1^{(k)}(\zeta) - \frac{k-2}{2k-1}\,z\,U_1^{(k-1)}(\zeta) + U_2^{(k-1)}(\zeta),\\
u_2^{(k)}(\zeta) &= \frac{1}{2}\,r\,V_1^{(k)}(\zeta) + V_3^{(k+1)}(\zeta),\\
u_z^{(k)}(\zeta) &= \frac{k-2}{2(2k-1)}\,z\,V_1^{(k)}(\zeta) + \frac{k+1}{2(2k-1)}\,r\,U_1^{(k-1)}(\zeta) + \frac{1}{2}\left(U_3^{(k)}(\zeta) - V_2^{(k)}(\zeta)\right).
\end{aligned}$$

With the relationships $u_r^{(k)} = \frac{1}{2}(u_1^{(k)} + u_2^{(k)})$ and $u_\varphi^{(k)} = \frac{1}{2}(u_2^{(k)} - u_1^{(k)})$, the representation (16) follows from (24). □

PROPOSITION 3. *Another solution form for the asymmetric velocity field in terms of three k-harmonically analytic functions vanishing at infinity is given by*

(25)
$$\begin{aligned}
2\left(u_z^{(k)}(\zeta) + i\,u_r^{(k)}(\zeta)\right) &= \left(\frac{k+2}{2k+1}\,z + i\frac{k-1}{2k+1}\,r\right) G_1^{(k)}(\zeta) + i\,G_2^{(k-1)}(\zeta) + G_3^{(k)}(\zeta),\\
2u_\varphi^{(k)}(\zeta) &= \operatorname{Im}\left[\frac{k+2}{2k+1}(z - i\,r)\,G_1^{(k)}(\zeta) - i\,G_2^{(k-1)}(\zeta) + G_3^{(k)}(\zeta)\right]
\end{aligned}$$

for $k \geq 1$, and the kth harmonic of the pressure is determined by

$$\wp^{(k)}(\zeta) = \mu\,\operatorname{Re} G_1^{(k)}(\zeta), \quad k \geq 1.$$

Proof. The proof is similar to that of Proposition 2. □

Without loss of generality, an asymmetric motion of a solid body of revolution, whose axis of revolution is determined by the z-axis, can be decomposed into two motions: (i) translation of the body along the x-axis with the constant velocity v_x ("x-translation"), and (ii) rotation of the body around the y-axis with the constant angular velocity ϖ_y ("y-rotation"). For these motions, the no-slip boundary conditions (2) for the velocity field $\mathbf{u}$ on the body's surface S take the form

$$x\text{-translation:} \quad \mathbf{u} = v_x\mathbf{i} \quad \text{on} \quad S, \tag{26}$$
$$y\text{-rotation:} \quad \mathbf{u} = [\varpi_y\,\mathbf{j} \times (x\,\mathbf{i} + z\,\mathbf{k})] \quad \text{on} \quad S, \tag{27}$$

and in both problems, $\mathbf{u}$ and $\wp$ vanish at infinity, i.e., satisfy (3).

For the components (u_r, u_φ, u_z) of the velocity field in the cylindrical coordinates (r, φ, z), the boundary conditions (26) and (27) are reformulated as

$$x\text{-translation:} \quad u_r = v_x\cos\varphi, \ \ u_\varphi = -v_x\sin\varphi, \ \ u_z = 0 \quad \text{on} \quad S, \tag{28}$$
$$y\text{-rotation:} \quad u_r = \varpi_y\,z\,\cos\varphi, \ \ u_\varphi = -\varpi_y\,z\,\sin\varphi, \ \ u_z = -\varpi_y\,r\,\cos\varphi \quad \text{on} \quad S. \tag{29}$$

It follows from (28) and (29) that the velocity field $\mathbf{u}$ has only a first harmonic with respect to the angular coordinate φ, and consequently, its components can be represented in the form (16) for $k=1$:

$$u_z^{(1)}(\zeta)+i\,u_r^{(1)}(\zeta)=\frac{1}{2}\left((2r+i\,z)\,G_1^{(0)}(\zeta)+i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)\right), \tag{30a}$$

$$u_\varphi^{(1)}(\zeta)=\frac{1}{2}\,\mathrm{Im}\left[-(r+i\,z)\,G_1^{(0)}(\zeta)-i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)\right]. \tag{30b}$$

Consequently, asymmetric problems of 3D Stokes flows due to motion of the solid body of revolution reduce to boundary-value problems for two 0-harmonically analytic functions and one 1-harmonically analytic function.

In the rz-plane, let $\mathcal{D}^+$ and $\mathcal{D}^-$ denote the inner and outer regions with respect to the cross-section of the finite body of revolution.[3] By $\mathcal{D}_0^+$ and $\mathcal{D}_0^-$, we will understand $\mathcal{D}^+$ with $r\geq 0$ and $\mathcal{D}^-$ with $r\geq 0$, respectively, i.e., the right parts of the corresponding regions. In general, $\mathcal{D}_0^+$ is a multiply connected region, e.g., a cross-section of two spheres. Let ℓ be the common boundary of $\mathcal{D}^+$ and $\mathcal{D}^-$, and let ℓ_+ and ℓ_- denote the parts of ℓ for $r\geq 0$ (right part) and $r\leq 0$ (left part), respectively, which, being symmetric with respect to the z-axis, are either closed curves or open curves with the endpoints lying on the z-axis. The contour of the body in the rz-plane is thus $\ell=\ell_+\bigcup\ell_-$. It is positively oriented, i.e., traversed in the counterclockwise direction, if $\mathcal{D}^+$ remains on the left side when one travels along ℓ in this direction.[4]

Problem I. *Given a complex-valued function $f_1(\zeta)$ and real-valued function $f_2(\zeta)$ on ℓ_+ such that $f_1\left(-\overline{\zeta}\right)=-\overline{f_1(\zeta)}$ and $f_2\left(-\overline{\zeta}\right)=f_2(\zeta)$, find 0-harmonically analytic functions $G_1^{(0)}(\zeta)$ and $G_2^{(0)}(\zeta)$ and a 1-harmonically analytic function $G_3^{(1)}(\zeta)$ in $\mathcal{D}_0^-$ that vanish at $|\zeta|\to\infty$ and satisfy the boundary conditions*

$$(2r+i\,z)\,G_1^{(0)}(\zeta)+i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)=f_1(\zeta),\quad \zeta\in\ell_+, \tag{31a}$$

$$\mathrm{Im}\left[-(r+i\,z)G_1^{(0)}(\zeta)-i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)\right]=f_2(\zeta),\quad \zeta\in\ell_+. \tag{31b}$$

For example, for the x-translation (28), we have $f_1(\zeta)=2iv_x$ and $f_2(\zeta)=-2v_x$, and for the y-rotation (29), we have $f_1(\zeta)=-2\varpi_y\overline{\zeta}$ and $f_2(\zeta)=-2\varpi_y z$.

Remark 1. With the symmetry condition (15) and with $f_1\left(-\overline{\zeta}\right)=-\overline{f_1(\zeta)}$ and $f_2\left(-\overline{\zeta}\right)=f_2(\zeta)$, the boundary conditions (31a) and (31b) are equivalent to the boundary conditions on $\ell=\ell_+\bigcup\ell_-$:

$$(2r+i\,z)\,G_1^{(0)}(\zeta)+i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)=f_1(\zeta),\quad \zeta\in\ell, \tag{32a}$$

$$\mathrm{Im}\left[-(r+i\,z)G_1^{(0)}(\zeta)-i\,G_2^{(0)}(\zeta)+G_3^{(1)}(\zeta)\right]=f_2(\zeta),\quad \zeta\in\ell. \tag{32b}$$

For details, see Remark 7 in [32]. This fact will be critical in reducing Problem I to integral equations.

Remark 2. For multiply connected $\mathcal{D}_0^-$, the functions $G_1^{(0)}(\zeta)$, $G_2^{(0)}(\zeta)$, and $G_3^{(1)}(\zeta)$ are continuous and single valued. Indeed, in this case, $G_1^{(0)}(\zeta)$, $G_2^{(0)}(\zeta)$, and $G_3^{(1)}(\zeta)$ may contain multivalued terms analogous to a complex logarithm, which change their values along a continuous closed path enclosing a branch point lying in $\mathcal{D}_0^+$; see [1, formula (32.22)] and [11, formulae (35.2), (36.4)]. However, the solution form (30a)–(30b) and the continuity of the velocity field and pressure imply that those terms vanish.

[3] We always assume that the z-axis is the body's axis of revolution.

[4] The orientation of a closed curve is always determined with respect to the corresponding inner region.

Proposition 4. *Problem* I *has a unique solution in the outer region* D_0^-.

Proof. The proposition is equivalent to the fact that (31a) and (31b) have only a zero homogeneous solution in the specified class of functions. Since the Stokes flow problem (1) and (3) with zero boundary condition $\mathbf{u}|_S = 0$ has only a zero solution (see Proposition 1), we have $u_r^{(1)} \equiv 0$, $u_\varphi^{(1)} \equiv 0$ and $u_z^{(1)} \equiv 0$ in $\mathcal{D}_0^-$, which in terms of $G_1^{(0)}(\zeta) = U_1^{(0)}(\zeta) + i\,V_1^{(1)}(\zeta)$, $G_2^{(0)}(\zeta) = U_2^{(0)}(\zeta) + i\,V_2^{(1)}(\zeta)$, and $G_3^{(1)}(\zeta) = U_3^{(1)}(\zeta) + i\,V_3^{(2)}(\zeta)$ with (30a)–(30b) reduces to

$$2r\,U_1^{(0)} - z\,V_1^{(1)} - V_2^{(1)} + U_3^{(1)} = 0, \tag{33a}$$

$$2r\,V_1^{(1)} + z\,U_1^{(0)} + U_2^{(0)} + V_3^{(2)} = 0, \tag{33b}$$

$$-\left(r\,V_1^{(1)} + z\,U_1^{(0)}\right) - U_2^{(0)} + V_3^{(2)} = 0 \tag{33c}$$

in the whole region $\mathcal{D}_0^-$. We should prove that the system (33a)–(33c) has only a zero solution. Indeed, applying the 1-harmonic operator to (33a) and using the first equation of (12) for $G_1^{(0)}(\zeta)$, we obtain $\frac{\partial}{\partial r}U_1^{(0)} = 0$ in $\mathcal{D}_0^-$. Subtracting (33c) from (33b) and applying the 0-harmonic operator to the resulting equation, we have $\frac{\partial}{\partial z}U_1^{(0)} = 0$ in $\mathcal{D}_0^-$. Consequently, we conclude that $U_1^{(0)}$ is a constant, which however, equals zero, since $G_1^{(0)}$ vanishes at infinity. With $U_1^{(0)} \equiv 0$, (12) implies that $V_1^{(1)} = b/r$, where b is a constant, which also equals zero because of the same reason, and thus, $G_1^{(0)} \equiv 0$ in $\mathcal{D}_0^-$. Substituting this result into (33b) and (33c), we obtain $U_2^{(0)} \equiv 0$ and $V_3^{(2)} \equiv 0$. Similarly, in this case, the only $G_2^{(0)}$ and $G_3^{(1)}$ that vanish at infinity are zero functions. □

Remark 3. If in Problem I the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are not required to vanish at infinity, then in this case, Problem I has the homogeneous solution $G_1^{(0)}(\zeta) = a$, $G_2^{(0)}(\zeta) = -a\left(z - \frac{i}{2}r\right)$, and $G_3^{(1)}(\zeta) = -\frac{3}{2}a\,r$, where a is an arbitrary constant. This fact can be established by modifying the proof of Proposition 4.

Further, we will need the following result.

Proposition 5 (auxiliary homogeneous boundary-value problem). *For* 0*-harmonically analytic functions* $G_1^{(0)}(\zeta)$ *and* $G_2^{(0)}(\zeta)$ *in the inner (multiply connected) region* $\mathcal{D}_0^+$*, the homogeneous boundary-value problem*

$$\left((2r + i\,z)\,G_1^{(0)}(\zeta) + i\,G_2^{(0)}(\zeta)\right)\Big|_{\ell_+} = 0 \tag{34}$$

has only zero solution.

Proof. The proof is analogous to that of Proposition 9(ii) in [32]. We assume that $\mathcal{D}_0^+$ is simply connected. In the case when $\mathcal{D}_0^+$ is multiply connected and consists of disjoint simply connected subregions $\mathcal{D}_j^+$, $1 \le j \le m$, the proof is conducted for $\mathcal{D}_j^+$ instead of $\mathcal{D}_0^+$. Let a function $\Phi(\zeta)$ be defined by

$$\Phi(\zeta) = r\,G_1^{(0)}(\zeta)\left((2r + i\,z)\,G_1^{(0)}(\zeta) + i\,G_2^{(0)}(\zeta)\right), \tag{35}$$

and let L be the positively oriented boundary of $\mathcal{D}_0^+$, which either is ℓ_+ if ℓ_+ is closed or consists of ℓ_+ and the segment of the z-axis connecting the endpoints of ℓ_+ if ℓ_+ is an open curve with the endpoints lying on the z-axis. It follows from (35) and Proposition 4(ii) in [32] that

$$\operatorname{Re}\frac{\partial\Phi}{\partial\bar{\zeta}} = \frac{1}{2}r\left(3\left[\operatorname{Re}G_1^{(0)}\right]^2 + \left[\operatorname{Im}G_1^{(0)}\right]^2\right).$$

Consequently, using Propositions 5 in [32], we obtain

(36)
$$\operatorname{Im}\left[\oint_L \Phi\, d\zeta\right] = \operatorname{Im}\left[2i \iint_{\mathcal{D}_0^+} \frac{\partial \Phi}{\partial \overline{\zeta}}\, drdz\right] = \iint_{\mathcal{D}_0^+} r\left(3\left[\operatorname{Re} G_1^{(0)}\right]^2 + \left[\operatorname{Im} G_1^{(0)}\right]^2\right) drdz.$$

However, since $\Phi = 0$ on the z-axis and ℓ_+, the integral in the right-hand side of the last equality vanishes, whence it follows that $G_1^{(0)}(\zeta) \equiv 0$ in $\mathcal{D}_0^+$. □

Problem I can be reduced to integral equations based on the generalized Cauchy integral formula

$$G^{(k)}(\zeta) = -\frac{1}{2\pi i}\oint_\ell G^{(k)}(\tau)\mathcal{W}^{(k)}(\zeta,\tau)\, d\tau, \qquad \zeta \in \operatorname{int}\mathcal{D}^-,$$

for k-harmonically analytic functions in the outer region $\mathcal{D}^-$ that vanish at infinity, where $\mathcal{W}^{(k)}(\zeta,\tau)$ is a generalized Cauchy kernel and ℓ is the positively oriented boundary with respect to $\mathcal{D}^+$; see Theorem 2 in the first part of this work [32].

Theorem 6 (two integral equations in the asymmetric case). *For the outer multiply connected region $\mathcal{D}_0^-$, Problem* I *reduces to two integral equations for determining boundary values of $G_1^{(0)}(\zeta) = U_1^{(0)}(\zeta) + i\, V_1^{(1)}(\zeta)$ and $U_3^{(1)}(\zeta) = \operatorname{Re} G_3^{(1)}(\zeta)$:*

(37)
$$\begin{aligned}
&\frac{1}{\pi i}\int_{\ell_+}\Bigg([z - z_1 - 2i(r - r_1)]\, G_1^{(0)}(\tau)\, \Omega_+^{(0)}(\zeta,\tau)\, \frac{d\tau}{\tau - \zeta} \\
&\qquad\qquad - [z - z_1 - 2i(r + r_1)]\, \overline{G_1^{(0)}(\tau)}\, \Omega_-^{(0)}(\zeta,\tau)\, \frac{d\overline{\tau}}{\overline{\tau} + \zeta}\Bigg) \\
&+ \frac{1}{\pi i}\int_{\ell_+}\Bigg(\left[\frac{1}{2} r_1 V_1^{(1)}(\tau) + i\, U_3^{(1)}(\tau)\right]\left(\Omega_+^{(0)}(\zeta,\tau) - \Omega_+^{(1)}(\zeta,\tau)\right)\frac{d\tau}{\tau - \zeta} \\
&\qquad\qquad - \left[\frac{1}{2} r_1 V_1^{(1)}(\tau) - i\, U_3^{(1)}(\tau)\right]\left(\Omega_-^{(0)}(\zeta,\tau) + \Omega_-^{(1)}(\zeta,\tau)\right)\frac{d\overline{\tau}}{\overline{\tau} + \zeta}\Bigg) = F_1(\zeta), \quad \zeta \in \ell,
\end{aligned}$$

(38)
$$U_3^{(1)}(\zeta) - \frac{i}{2} r\, V_1^{(1)}(\zeta) + \frac{1}{\pi i}\oint_\ell \left(U_3^{(1)}(\tau) - \frac{i}{2} r_1\, V_1^{(1)}(\tau)\right)\mathcal{W}^{(1)}(\zeta,\tau)\, d\tau = F_2(\zeta), \quad \zeta \in \ell,$$

where $\zeta = r + i\, z$, $\tau = r_1 + i\, z_1$, the functions $\Omega_+^{(k)}(\zeta,\tau)$ and $\Omega_-^{(k)}(\zeta,\tau)$ are determined by (25a) *and* (25b) *in* [32], *and*

(39)
$$F_1(\zeta) = i\, f_1(\zeta) + \frac{1}{2\pi i}\oint_\ell (2i\, f_1(\tau) + f_3(\tau))\, \mathcal{W}^{(0)}(\zeta,\tau)\, d\tau + \frac{1}{2\pi}\oint_\ell i\, f_3(\tau)\mathcal{W}^{(1)}(\zeta,\tau)\, d\tau,$$

(40)
$$F_2(\zeta) = -\frac{i}{2} f_3(\zeta) - \frac{1}{2\pi i}\int_\ell i\, f_3(\tau)\mathcal{W}^{(1)}(\zeta,\tau)\, d\tau,$$

$$f_3(\zeta) = \operatorname{Im}[f_1(\zeta)] + f_2(\zeta).$$

Proof. As in the axially symmetric case [32, Theorem 10], the derivation of the integral equations follows Muskhelishvili's approach [11, 12] used for reducing 2D problems of an elastic medium to integral equations.

According to Remark 1, Problem I for $\mathcal{D}_0^-$ is equivalent to Problem I for $\mathcal{D}^-$. Necessary and sufficient conditions for the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ to be k-harmonically analytic in $\mathcal{D}^-$ for $k = 0$, 0, and 1, respectively, and vanishing at infinity follow from the corresponding generalized Sokhotski–Plemelj formulae[5] and are given by

$$G_1^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell G_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = 0, \quad \zeta \in \ell, \tag{41a}$$

$$G_2^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell G_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = 0, \quad \zeta \in \ell, \tag{41b}$$

$$G_3^{(1)}(\zeta) + \frac{1}{\pi i}\oint_\ell G_3^{(1)}(\tau)\mathcal{W}^{(1)}(\zeta,\tau)\,d\tau = 0, \quad \zeta \in \ell. \tag{41c}$$

Expressing the boundary value of $G_2^{(0)}$ from the boundary condition (32a),

$$G_2^{(0)}(\zeta) = -i\left(f_1(\zeta) - (2r + i\,z)\,G_1^{(0)}(\zeta) - G_3^{(1)}(\zeta)\right), \quad \zeta \in \ell, \tag{42}$$

and substituting (42) into (41b), we have

$$\begin{aligned}(2ir - z)\,G_1^{(0)}(\zeta) + i\,G_3^{(1)}(\zeta) + \frac{1}{\pi i}\oint_\ell \left((2ir_1 - z_1)G_1^{(0)}(\tau) + i\,G_3^{(1)}(\tau)\right)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau \\ = i\,f_1(\zeta) + \frac{1}{\pi i}\oint_\ell i\,f_1(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \ell.\end{aligned} \tag{43}$$

Similarly, expressing the boundary value of $V_3^{(2)}(\zeta)$ from the sum of (32b) and the imaginary part of (32a),

$$V_3^{(2)}(\zeta) = \frac{1}{2}\left(\mathrm{Im}[f_1(\zeta)] + f_2(\zeta) - r\,V_1^{(1)}(\zeta)\right), \quad \zeta \in \ell, \tag{44}$$

and substituting (44) into (41c), we obtain (38). Finally, substituting (44) into (43) and subtracting the combination $(2ir - z)\cdot(41\text{a}) + i\cdot(38)$ from (43), we obtain (37).

Now we need to show that if $\widetilde{G}_1^{(0)}(\zeta)$ and $\widetilde{U}_3^{(1)}(\zeta)$ solve (37) and (38), then $\widetilde{G}_1^{(0)}(\zeta)$, $\widetilde{G}_2^{(0)}(\zeta)$, determined by (42), and $\widetilde{G}_3^{(1)}(\zeta)$ satisfy (41a), (41b), and (41c), respectively.

With $\widetilde{V}_1^{(1)}$, the imaginary part $\widetilde{V}_3^{(2)}(\zeta)$ is determined by (44), and (38) can be restated as (41c) for $\widetilde{G}_3^{(1)}(\zeta) = \widetilde{U}_3^{(1)}(\zeta) + i\,\widetilde{V}_3^{(2)}(\zeta)$. Thus, $\widetilde{G}_3^{(1)}(\zeta)$ satisfies (41c). Similarly, with (44), (37) can be rewritten in terms of $\widetilde{G}_3^{(1)}(\zeta)$, and the combination $(37) + i\cdot(41\text{c})$ reduces to

$$\begin{aligned}&\frac{1}{\pi i}\left(\oint_\ell (2ir_1 - z_1)\widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau - (2ir - z)\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau\right) \\ &\qquad + i\,\widetilde{G}_3^{(1)}(\zeta) + \frac{1}{\pi i}\oint_\ell i\,\widetilde{G}_3^{(1)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau \\ &\qquad = i\,f_1(\zeta) + \frac{1}{\pi i}\oint_\ell i\,f_1(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \ell.\end{aligned} \tag{45}$$

[5] These formulae are derived similarly to the Sokhotski–Plemelj formulae for ordinary analytic functions; see [1, formula (31.13)] and [7, formula (4.8)].

Now adding $(2ir - z)\widetilde{G}_1^{(0)}(\zeta)$ to the right-hand and left-hand sides of (45) and using (42), we rewrite (45) in the form

$$\begin{aligned} &-(2ir - z)\left(\widetilde{G}_1^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau\right) \\ &\qquad + \widetilde{G}_2^{(0)}(\zeta) + \frac{1}{\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau = 0, \quad \zeta \in \ell. \end{aligned} \tag{46}$$

Let $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$ be determined by the generalized Cauchy-type integrals in the region $\mathcal{D}^+$ excluding its boundary ℓ:

$$\Phi^+(\zeta) = \frac{1}{2\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \operatorname{int}\mathcal{D}^+,$$
$$\Psi^+(\zeta) = \frac{1}{2\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \operatorname{int}\mathcal{D}^+.$$

These functions are 0-harmonically analytic in $\operatorname{int}\mathcal{D}^+$, since $\mathcal{W}^{(0)}(\zeta,\tau)$ satisfies (14) for $k = 0$ with respect to ζ. Then when ζ approaches ℓ from within $\mathcal{D}^+$, the boundary values of $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$ on ℓ are determined by the corresponding generalized Sokhotski–Plemelj formula

$$\Phi^+(\zeta) = \frac{1}{2}\widetilde{G}_1^{(0)}(\zeta) + \frac{1}{2\pi i}\oint_\ell \widetilde{G}_1^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \ell, \tag{47a}$$

$$\Psi^+(\zeta) = \frac{1}{2}\widetilde{G}_2^{(0)}(\zeta) + \frac{1}{2\pi i}\oint_\ell \widetilde{G}_2^{(0)}(\tau)\mathcal{W}^{(0)}(\zeta,\tau)\,d\tau, \quad \zeta \in \ell, \tag{47b}$$

and the relationship (46) reduces to

$$(2r + i\,z)\,\Phi^+(\zeta) + i\,\Psi^+(\zeta) = 0, \quad \zeta \in \ell,$$

which, in view of the symmetry condition (15),[6] is the auxiliary homogeneous boundary-value problem (34) for the 0-harmonically analytic functions $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$ in $\mathcal{D}_0^+$. According to Proposition 5, the only solution to this problem is $\Phi^+(\zeta) \equiv 0$ and $\Psi^+(\zeta) \equiv 0$. Consequently, (47a) and (47b) imply that $\widetilde{G}_1^{(0)}(\zeta)$ and $\widetilde{G}_2^{(0)}(\zeta)$ satisfy (41a) and (41b), respectively, and thus are the boundary values of 0-harmonically analytic functions in $\mathcal{D}^-$. □

Remark 4. The kernels of the integral equation (37) have only logarithmic singularity because of the functions $\Omega_-^{(0)}(\zeta,\tau)$ and $\Omega_-^{(1)}(\zeta,\tau)$ (see Remark 2 in [32]). Indeed,

$$\lim_{\tau\to\zeta} \frac{z - z_1 - 2i(r - r_1)}{\tau - \zeta} = \frac{i}{2}\left(3 + \mathrm{e}^{-2i \lim\limits_{\tau\to\zeta} \arg[\tau-\zeta]}\right),$$

which is obtained by setting $\tau = \zeta + \rho\,\mathrm{e}^{i\beta}$ and passing $\rho \to 0$ (also, if $\zeta = \zeta(t)$ is a parameterization of smooth ℓ_+, then $\lim_{\tau\to\zeta}\arg[\tau - \zeta] = \arg[\zeta'(t)]$), and

$$\lim_{\tau\to\zeta} \frac{\Omega_+^{(0)}(\zeta,\tau) - \Omega_+^{(1)}(\zeta,\tau)}{\tau - \zeta} = 0.$$

[6] The functions $\Phi^+(\zeta)$ and $\Psi^+(\zeta)$, being determined by the generalized Cauchy-type integrals, satisfy the symmetry condition (15).

Remark 5. Let ℓ_+ be symmetric with respect to the r-axis. If $f_1\left(\overline{\zeta}\right) = \overline{f_1(\zeta)}$ and $f_2\left(\overline{\zeta}\right) = -f_2(\zeta)$, then $G_3^{(1)}\left(\overline{\zeta}\right) = \overline{G_3^{(1)}(\zeta)}$, and (38) reduces to

$$U_3^{(1)}(\zeta) + \mathrm{Re}\left[\frac{1}{\pi i}\oint_\ell \left(U_3^{(1)}(\tau) - \frac{i}{2} r_1\, V_1^{(1)}(\tau)\right)\mathcal{W}^{(1)}(\zeta,\tau)\,d\tau\right] = \mathrm{Re}[F_2(\zeta)], \quad \zeta\in\ell,$$

and if $f_1\left(\overline{\zeta}\right) = -\overline{f_1(\zeta)}$ and $f_2\left(\overline{\zeta}\right) = f_2(\zeta)$, then $G_3^{(1)}\left(\overline{\zeta}\right) = -\overline{G_3^{(1)}(\zeta)}$, and (38) reduces to

$$-\frac{1}{2} r\, V_1^{(1)}(\zeta) + \mathrm{Im}\left[\frac{1}{\pi i}\oint_\ell \left(U_3^{(1)}(\tau) - \frac{i}{2} r_1\, V_1^{(1)}(\tau)\right)\mathcal{W}^{(1)}(\zeta,\tau)\,d\tau\right] = \mathrm{Im}[F_2(\zeta)], \ \ \zeta\in\ell.$$

These equations follow from Proposition 10 (see the appendix) applied to (41c) with (44).

The singular integral equation (38) can be reduced to an integral equation with a logarithmic singularity based on Proposition 11 (see the appendix) and the fact that (38) is (41c) with (44).

Now we represent the resisting force and torque, exerted on a solid body of revolution in the x-translation (26) and y-rotation (27), respectively, in terms of the functions $G_1^{(0)}$ and $G_2^{(0)}$, entering the solution form (30a)–(30b).

PROPOSITION 7 (resisting force in the asymmetric translation). *For the Stokes flow due to the x-translation* (28) *of the body,*[7] *let the velocity field be represented by* (30a)–(30b). *The resisting (drag) force, exerted on the body by the fluid, can be represented in two equivalent forms,*

$$F_x = 2\pi\mu\ \mathrm{Re}\left[\int_{\ell_+} r\, G_1^{(0)}(\zeta)\,d\zeta\right], \tag{48a}$$

$$F_x = -4\pi\mu \lim_{z\to\infty}\left(z^2\ \mathrm{Re}\, G_1^{(0)}(r,z)\Big|_{r=0}\right), \tag{48b}$$

where ℓ_+ in (48a) *is positively oriented with respect to $\mathcal{D}_0^+$.*

Proof. We first prove the formula (48a).

For the x-translation (28), the resulting force, exerted on the body of revolution, is the integral over the body's surface S and has the component in the direction $\mathbf{i}$ only:

$$F_x = \iint_S (\mathbf{i}\cdot\mathbf{P}_n)\,dS, \qquad \mathbf{P}_n = 2\mu\,\frac{\partial\mathbf{u}}{\partial n} + \mu\,[\mathfrak{n}\times\mathrm{curl}\,\mathbf{u}] - \wp\,\mathfrak{n}, \tag{49}$$

where $\mathfrak{n} = n_r\,\mathbf{e}_r + n_z\,\mathbf{k}$ is the outer normal to the body's surface with $n_r = \frac{\partial r}{\partial n}$ and $n_z = \frac{\partial z}{\partial n}$; see [9].

Using the representations (30a)–(30b) and (17) for $k=1$ that correspond to the boundary conditions (28), we have

$$\begin{aligned}
u_r(r,\varphi,z) &= \frac{1}{2}\left(z\,U_1^{(0)}(r,z) + 2r\,V_1^{(1)}(r,z) + U_2^{(0)}(r,z) + V_3^{(2)}(r,z)\right)\cos\varphi,\\
u_\varphi(r,\varphi,z) &= \frac{1}{2}\left(-z\,U_1^{(0)}(r,z) - r\,V_1^{(1)}(r,z) - U_2^{(0)}(r,z) + V_3^{(2)}(r,z)\right)\sin\varphi,\\
u_z(r,\varphi,z) &= \frac{1}{2}\left(2r\,U_1^{(0)}(r,z) - z\,V_1^{(1)}(r,z) - V_2^{(1)}(r,z) + U_3^{(1)}(r,z)\right)\cos\varphi,\\
\wp(r,\varphi,z) &= \mu\,V_1^{(1)}(r,z)\cos\varphi.
\end{aligned} \tag{50}$$

[7] The z-axis is the body's axis of revolution.

Let (s,φ,n) be a characteristic coordinate system with the right-hand orthogonal basis $(\mathfrak{s},\mathbf{e}_\varphi,\mathfrak{n})$, in which s has *negative* orientation. Then with the relationships

$$\frac{\partial r}{\partial s}=\frac{\partial z}{\partial n}, \qquad \frac{\partial r}{\partial n}=-\frac{\partial z}{\partial s},$$

the system (12) with $k=0$, defining the functions $G_1^{(0)}=U_1^{(0)}+i\,V_1^{(1)}$ and $G_2^{(0)}=U_2^{(0)}+i\,V_2^{(1)}$, can be represented by

$$\frac{\partial}{\partial s}U_j^{(0)}=\frac{1}{r}\frac{\partial}{\partial n}\left(rV_j^{(1)}\right), \qquad \frac{\partial}{\partial n}U_j^{(0)}=-\frac{1}{r}\frac{\partial}{\partial s}\left(rV_j^{(1)}\right), \quad j=1,2, \tag{51}$$

and (12) with $k=1$, defining the function $G_3^{(1)}=U_3^{(1)}+i\,V_3^{(2)}$, takes the form

$$\begin{aligned}\frac{\partial}{\partial s}U_3^{(1)}-\frac{1}{r}\frac{\partial r}{\partial s}U_3^{(1)}&=\frac{1}{r^2}\frac{\partial}{\partial n}\left(r^2V_3^{(2)}\right),\\ \frac{\partial}{\partial n}U_3^{(1)}-\frac{1}{r}\frac{\partial r}{\partial n}U_3^{(1)}&=-\frac{1}{r^2}\frac{\partial}{\partial s}\left(r^2V_3^{(2)}\right).\end{aligned} \tag{52}$$

In the cylindrical coordinates, $dS=r\,ds\,d\varphi$, where ds is the differential of the curve length (s is the same variable as in the system (s,φ,n)). Substituting (50) into (49) and also using (12) along with (51) and (52), we obtain

$$\begin{aligned}\int_0^{2\pi}(\mathbf{i}\cdot\mathbf{P}_n)\,d\varphi=\frac{\pi\mu}{2r}\Bigg(&4r\left(-\frac{\partial r}{\partial s}U_1^{(0)}+\frac{\partial z}{\partial s}V_1^{(1)}\right)\\ &+\frac{\partial}{\partial s}\left[r\left(5r\,U_1^{(0)}-3z\,V_1^{(1)}-3\,V_2^{(1)}+U_3^{(1)}\right)\right]\Bigg),\end{aligned}$$

and consequently, F_x in (49) reduces to

$$F_x=\int_{\ell_+}\left[\int_0^{2\pi}(\mathbf{i}\cdot\mathbf{P}_n)\,d\varphi\right]r\,ds=2\pi\mu\int_{\ell_+}r\left(-\frac{\partial r}{\partial s}U_1^{(0)}+\frac{\partial z}{\partial s}V_1^{(1)}\right)ds, \tag{53}$$

under the condition that

$$\int_{\ell_+}\frac{\partial}{\partial s}\left[r\left(5r\,U_1^{(0)}-3z\,V_1^{(1)}-3V_2^{(1)}+U_3^{(1)}\right)\right]ds=0,$$

which obviously holds, since the functions $G_1^{(0)}(\zeta)$, $G_2^{(0)}(\zeta)$, and $G_3^{(1)}(\zeta)$ are continuous in $\mathcal{D}_0^-$ (see Remark 2) and since ℓ_+ either is a closed curve or has the endpoints on the z-axis.

Finally, since ℓ_+ in (53) has positive orientation, $dr=-\frac{\partial r}{\partial s}ds$ and $dz=-\frac{\partial z}{\partial s}ds$ on ℓ_+, and thus, with these relationships, (53) can be represented in the form of (48a).

The formula (48b) follows from (48a) and Proposition 6 in [32]. □

PROPOSITION 8 (resisting torque in the asymmetric rotation). *For the Stokes flow due to the y-rotation* (29) *of the body,*[8] *let the velocity field be represented by* (30a)–(30b). *The resisting torque, exerted on the body by the fluid, can be represented in two equivalent forms,*

$$T_y=2\pi\mu\,\mathrm{Re}\left[\int_{\ell_+}r\left((2z-ir)\,G_1^{(0)}(\zeta)+G_2^{(0)}(\zeta)\right)d\zeta\right], \tag{54a}$$

$$T_y=4\pi\mu\left\{2\lim_{r\to\infty}\left(r^3\,\mathrm{Re}\,G_1^{(0)}(r,z)\Big|_{z=0}\right)-\lim_{z\to\infty}\left(z^2\,\mathrm{Re}\,G_2^{(0)}(r,z)\Big|_{r=0}\right)\right\}, \tag{54b}$$

where ℓ_+ in (54a) *is positively oriented with respect to $\mathcal{D}_0^+$.*

[8] The z-axis is the body's axis of revolution.

Proof. We first prove the formula (54a).

For the y-rotation (29), the resulting torque, exerted on the body of revolution, has the component in the direction $\mathbf{j}$ only:

$$T_y = \iint_S (\mathbf{j}\cdot[\mathfrak{r}\times\mathbf{P}_n])\,dS = \iint_S ([\mathbf{j}\times\mathfrak{r}]\cdot\mathbf{P}_n)\,dS, \tag{55}$$

where $\mathbf{P}_n$ is defined as in (49), $\mathfrak{r} = r\,\mathbf{e}_r + z\,\mathbf{k}$ is the radius vector, and $\mathfrak{n}$ is the outer normal to the body's surface as in (49).

As in the proof of Proposition 7, we introduce a characteristic coordinate system (s,φ,n) with the right-hand orthogonal basis $(\mathfrak{s},\mathbf{e}_\varphi,\mathfrak{n})$ and also have $dS = r\,ds\,d\varphi$ in the cylindrical coordinates with the length differential ds. For the boundary conditions (29), the velocity field can also be represented in the form (50).

Substituting (50) into (55) and using (12) along with (51) and (52), we obtain

$$\begin{aligned}
&\int_0^{2\pi} ([\mathbf{j}\times\mathfrak{r}]\cdot\mathbf{P}_n)\,d\varphi \\
&= \frac{\pi\mu}{2r}\left(-4r\left[r\left(\frac{\partial z}{\partial s}U_1^{(0)} + \frac{\partial r}{\partial s}V_1^{(1)}\right) + 2z\left(\frac{\partial r}{\partial s}U_1^{(0)} - \frac{\partial z}{\partial s}V_1^{(1)}\right) + \frac{\partial r}{\partial s}U_2^{(0)} - \frac{\partial z}{\partial s}V_1^{(1)}\right]\right. \\
&\qquad \left. + \frac{\partial}{\partial s}\left[r\left(7rz\,U_1^{(0)} + (4r^2-3z^2)V_1^{(1)} + 2r\,U_2^{(0)} - 3z\,V_2^{(1)} + z\,U_3^{(1)} + 2r\,V_3^{(2)}\right)\right]\right),
\end{aligned}$$

and consequently, T_y in (55) reduces to

$$\begin{aligned}
T_y = -2\pi\mu\int_{\ell_+} r\left[r\left(\frac{\partial z}{\partial s}U_1^{(0)} + \frac{\partial r}{\partial s}V_1^{(1)}\right) + 2z\left(\frac{\partial r}{\partial s}U_1^{(0)} - \frac{\partial z}{\partial s}V_1^{(1)}\right)\right. \\
\left. + \frac{\partial r}{\partial s}U_2^{(0)} - \frac{\partial z}{\partial s}V_1^{(1)}\right] ds,
\end{aligned} \tag{56}$$

under the condition that

$$\int_{\ell_+} \frac{\partial}{\partial s}\left[r\left(7rz\,U_1^{(0)} + (4r^2-3z^2)V_1^{(1)} + 2r\,U_2^{(0)} - 3z\,V_2^{(1)} + z\,U_3^{(1)} + 2r\,V_3^{(2)}\right)\right] ds = 0,$$

which holds, since $G_1^{(0)}(\zeta)$, $G_2^{(0)}(\zeta)$, and $G_3^{(1)}(\zeta)$ are continuous in $\mathcal{D}_0^-$ (see Remark 2) and since ℓ_+ either is closed or has the endpoints on the z-axis. As in the proof of Proposition 7, $dr = -\frac{\partial r}{\partial s}ds$ and $dz = -\frac{\partial z}{\partial s}ds$ on ℓ_+, which is positively oriented, and consequently, (56) can be represented in the form of (54a).

The formula (54b) is obtained from (54a). Indeed, it follows from Proposition 6 in [32] that $\operatorname{Re}[\int_{\ell_+} r\,G_2^{(0)}(\zeta)\,d\zeta] = -2\lim_{z\to\infty}(z^2\operatorname{Re}G_2^{(0)}(r,z)|_{r=0})$. Also, as in Proposition 6 in [32], it can be shown that

$$\operatorname{Re}\left[\int_{\ell_+} r\,(2z - ir)\,G_1^{(0)}(\zeta)\,d\zeta\right] = 4\lim_{r\to\infty}\left(r^3\operatorname{Re}G_1^{(0)}(r,z)\Big|_{z=0}\right),$$

and formula (54b) follows. □

3. Exact solutions to asymmetric Stokes flow problems. As an illustration for the developed framework, we solve asymmetric 3D Stokes flow problems for the x-translation (28) and y-rotation (29) of a solid sphere and solid prolate and oblate spheroids using the solution form (30a)–(30b) and series representations for k-harmonically analytic functions for corresponding regions. We also solve the integral equations (37) and (38) for the x-translation of solid bispheroids (two separate spheroids of equal size and having the same axis of revolution) and for the y-rotation of a solid torus of elliptical cross-section for various values of a geometrical parameter.

Example 1 (asymmetric translation and rotation of a solid sphere). Let (R, ϑ, φ) be the spherical coordinates related to the cylindrical coordinates in the ordinary way, and let a solid sphere be centered at the origin and have radius c.

(i) For the x-translation of the sphere, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ in (30a)–(30b) are determined in the region $R \geq c$ by

$$G_1^{(0)}(R,\vartheta) = \frac{3v_x c}{2}\, R^{-2}\, \mathrm{e}^{i\vartheta}, \quad G_2^{(0)}(R,\vartheta) = \frac{v_x c^3}{8}\, R^{-3}\left(1 + 3\,\mathrm{e}^{2i\vartheta}\right), \quad G_3^{(1)}(R,\vartheta) = -\frac{3v_x c^3}{4}\, R^{-3}\, \mathrm{e}^{i\vartheta}\, \sin\vartheta, \tag{57}$$

and the drag force $F_x = -6\pi\mu c v_x$ follows from (48b) and (57).

(ii) For the y-rotation of the sphere, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ in (30a)–(30b) are determined in the region $R \geq c$ by

$$G_1^{(0)}(R,\vartheta) \equiv 0, \qquad G_2^{(0)}(R,\vartheta) = 2\varpi_y c^3\, R^{-2}\, \mathrm{e}^{i\vartheta}, \qquad G_3^{(1)}(R,\vartheta) \equiv 0, \tag{58}$$

and the resisting torque $T_y = -8\pi\mu c^3 \varpi_y$ follows from (54b) and (58).

Detail. For the components $u_1^{(1)}$ and $u_2^{(1)}$, related to $u_r^{(1)}$ and $u_\varphi^{(1)}$ by (6), and $u_z^{(1)}$, the boundary conditions (28) and (29) for the sphere take the form

$$x\text{-translation:}\quad u_1^{(1)}\Big|_{R=c} = 2v_x,\quad u_2^{(1)}\Big|_{R=c} = 0,\quad u_z^{(1)}\Big|_{R=c} = 0, \tag{59}$$

$$y\text{-rotation:}\quad u_1^{(1)}\Big|_{R=c} = 2\varpi_y\, c\,\cos\vartheta,\quad u_2^{(1)}\Big|_{R=c} = 0,\quad u_z^{(1)}\Big|_{R=c} = -\varpi_y\, c\,\sin\vartheta. \tag{60}$$

We seek to find solutions that satisfy (59) and (60) in the form (30a)–(30b). For the region exterior to a sphere, the k-harmonically analytic functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are represented in the spherical coordinates by the series (15) in [32] for $k = 0$, 0, and 1, respectively:

$$\begin{aligned} G_1^{(0)}(R,\vartheta) &= \sum_{n=1}^{\infty} A_n R^{-n-1}\left\{n\,\mathrm{P}_n(\cos\vartheta) - i\,\mathrm{P}_n^{(1)}(\cos\vartheta)\right\},\\ G_2^{(0)}(R,\vartheta) &= \sum_{n=1}^{\infty} B_n R^{-n-1}\left\{n\,\mathrm{P}_n(\cos\vartheta) - i\,\mathrm{P}_n^{(1)}(\cos\vartheta)\right\},\\ G_3^{(1)}(R,\vartheta) &= \sum_{n=1}^{\infty} C_n R^{-n-2}\left\{n\,\mathrm{P}_{n+1}^{(1)}(\cos\vartheta) - i\,\mathrm{P}_{n+1}^{(2)}(\cos\vartheta)\right\}. \end{aligned} \tag{61}$$

From (59), (60), and (30a)–(30b), along with (61), we obtain second-order difference equations for the coefficients A_n, B_n, and C_n:

$$x\text{-translation:}\quad \frac{4}{3}c^{-1}A_1 = 2v_x,\quad \frac{13}{5}A_2 + B_1 = 0,\quad \frac{7}{10}A_2 + \frac{1}{2}B_1 = 0, \tag{62}$$

$$y\text{-rotation:}\quad \frac{4}{3}c^{-1}A_1 = 0,\quad \frac{13}{5}A_2 + B_1 = 2\varpi_y c^3,\quad \frac{7}{10}A_2 + \frac{1}{2}B_1 = \varpi_y\, c^3, \tag{63}$$

and (for the x-translation and y-rotation)

$$\text{(64a)} \qquad -c^2\frac{n(n-1)}{2(2n-1)}A_{n-1}+\frac{(n+1)(5n+8)}{2(2n+3)}A_{n+1}+n\,B_n=0, \qquad n\geq 2,$$

$$\text{(64b)} \qquad \frac{c^2}{2(2n-1)}A_{n-1}-\frac{1}{2(2n+3)}A_{n+1}-C_{n-1}=0, \qquad n\geq 2,$$

$$\text{(64c)} \quad -c^2\frac{(n-1)}{2(2n-1)}A_{n-1}+\frac{(3n+4)}{2(2n+3)}A_{n+1}+\frac{1}{2}\,B_n+\frac{(n-1)}{2}C_{n-1}=0, \quad n\geq 2.$$

The combination $2n\cdot$(64c) $+\,n(n-1)\cdot$(64b) $-$ (64a) reduces to having $A_{n+1}=0$ for $n\geq 2$, and consequently, it follows from (64a)–(64c) for $n\geq 3$ that $B_n=0$ and $C_{n-1}=0$ for $n\geq 3$.

Solving (62) and (64a)–(64c) for $n=2$, we obtain $A_1=3\,c\,v_x/2$, $A_2=0$, $B_1=0$, $B_2=c^3v_x/4$, and $C_1=c^3v_x/4$, which results in (57); similarly, solving (63) and (64a)–(64c) for $n=2$, we have $A_1=0$, $A_2=0$, $B_1=2\varpi_y c^3$, $B_2=0$, and $C_1=0$, which corresponds to the solution (58).

Example 2 (asymmetric translation and rotation of a solid prolate spheroid). Let (ξ,η,φ) be the prolate spheroidal coordinates related to the cylindrical coordinates by

$$\text{(65)} \qquad r=c\,\sinh\xi\,\sin\eta, \quad z=c\,\cosh\xi\,\cos\eta, \qquad \xi\in[0,\infty), \quad \eta\in[0,\pi],$$

where the angular coordinate $\varphi\in[0,2\pi)$ coincides with the one in (r,φ,z) and c is a metric parameter. In (ξ,η,φ), a solid prolate spheroid with the z-axis of revolution is determined by fixing the coordinate ξ, i.e., $\xi=\xi_0$. For the x-translation (28) and y-rotation (29), let the velocity field be represented by (30a)–(30b).

(i) For the x-translation, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are determined in the region $\xi\geq\xi_0$ by

(66)

$$\begin{aligned}
G_1^{(0)}(\xi,\eta)&=\frac{q_1}{c}\,\frac{(\cos\eta+i\,\coth\xi\sin\eta)}{\cosh^2\xi-\cos^2\eta},\\
G_2^{(0)}(\xi,\eta)&=\frac{q_1}{2}\left(\cosh^2\xi_0-3\right)\left(\frac{\sinh[2\xi]+i\,\sin[2\eta]}{2\sinh\xi\left(\cosh^2\xi-\cos^2\eta\right)}-\ln\left(\coth[\xi/2]\right)\right),\\
G_3^{(1)}(\xi,\eta)&=-\frac{q_1}{2}\,\sinh^2\xi_0\,\frac{\sin\eta}{\sinh\xi}\,\frac{(\cos\eta+i\,\coth\xi\sin\eta)}{\cosh^2\xi-\cos^2\eta},
\end{aligned}$$

where $q_1=4v_x\left/\left(\cosh\xi_0-(\cosh^2\xi_0-3)\ln\left(\coth[\xi_0/2]\right)\right)\right.$. The drag force $F_x=-4\pi\mu c q_1$ follows from (48b) and (66).

(ii) For the y-rotation, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are determined in the region $\xi\geq\xi_0$ by

(67)

$$\begin{aligned}
G_1^{(0)}(\xi,\eta)&=\frac{4q_2}{c}\left(\frac{\sinh[2\xi]+i\,\sin[2\eta]}{2\sinh\xi\left(\cosh^2\xi-\cos^2\eta\right)}-\ln\left(\coth[\xi/2]\right)\right),\\
G_2^{(0)}(\xi,\eta)&=q_2\left(\cosh^2\xi_0+3\right)\left(2\mathrm{Q}_1(\cosh\xi)\cos\eta-i\,\mathrm{Q}_1^{(1)}(\cosh\xi)\sin\eta\right)\\
&\qquad+2q_2\left(\cosh^2\xi_0-3\right)\frac{(\cos\eta+i\,\coth\xi\sin\eta)}{\cosh^2\xi-\cos^2\eta},\\
G_3^{(1)}(\xi,\eta)&=-q_2\,\sinh^2\xi_0\sin\eta\left(3\mathrm{Q}_1^{(1)}(\cosh\xi)+\frac{\sinh[2\xi]+i\,\sin[2\eta]}{\sinh^2\xi\left(\cosh^2\xi-\cos^2\eta\right)}\right),
\end{aligned}$$

where $q_2 = -\varpi_y c \big/ \left(\cosh\xi_0 - \left(\cosh^2\xi_0 + 1\right)\ln\left(\coth[\xi_0/2]\right)\right)$, and

$$\mathrm{Q}_1(\cosh\xi) = \cosh\xi\,\ln\left(\coth[\xi/2]\right) - 1,$$
$$\mathrm{Q}_1^{(1)}(\cosh\xi) = \sinh\xi\,\ln\left(\coth[\xi/2]\right) - \coth\xi.$$

The resisting torque $T_y = -16\pi\mu c^2 q_2 \cosh[2\xi_0]/3$ follows from (54b) and (67).

Detail. In terms of the components $u_1^{(1)}$, $u_2^{(1)}$, and $u_z^{(1)}$ (see (6)), the boundary conditions (28) and (29) for the prolate spheroid take the form

$$\text{(68a)}\qquad x\text{-translation:}\quad u_1^{(1)}\Big|_{\xi=\xi_0} = 2v_x,\quad u_2^{(1)}\Big|_{\xi=\xi_0} = 0,\quad u_z^{(1)}\Big|_{\xi=\xi_0} = 0,$$

$$\text{(68b)}\qquad \begin{aligned} y\text{-rotation:}\quad & u_1^{(1)}\Big|_{\xi=\xi_0} = 2\varpi_y\, c\,\cosh\xi_0\cos\eta,\quad u_2^{(1)}\Big|_{\xi=\xi_0} = 0,\\ & u_z^{(1)}\Big|_{\xi=\xi_0} = -\varpi_y\, c\,\sinh\xi_0\sin\eta. \end{aligned}$$

For the region exterior to the prolate spheroid, $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are represented in the prolate spheroidal coordinates (65) by the series (20) in [32] for $k = 0$, 0, and 1, respectively:

(69)
$$\begin{aligned} G_1^{(0)}(\xi,\eta) &= \sum_{n=1}^{\infty} A_n \left\{ n(n+1)\,\mathrm{Q}_n(\cosh\xi)\,\mathrm{P}_n(\cos\eta) + i\,\mathrm{Q}_n^{(1)}(\cosh\xi)\,\mathrm{P}_n^{(1)}(\cos\eta)\right\},\\ G_2^{(0)}(\xi,\eta) &= \sum_{n=1}^{\infty} B_n \left\{ n(n+1)\,\mathrm{Q}_n(\cosh\xi)\,\mathrm{P}_n(\cos\eta) + i\,\mathrm{Q}_n^{(1)}(\cosh\xi)\,\mathrm{P}_n^{(1)}(\cos\eta)\right\},\\ G_3^{(1)}(\xi,\eta) &= \sum_{n=2}^{\infty} C_n \left\{ (n-1)(n+2)\,\mathrm{Q}_n^{(1)}(\cosh\xi)\,\mathrm{P}_n^{(1)}(\cos\eta)\right.\\ &\qquad\qquad \left. + i\,\mathrm{Q}_n^{(2)}(\cosh\xi)\,\mathrm{P}_n^{(2)}(\cos\eta)\right\}. \end{aligned}$$

Substituting (30a)–(30b) with (69) into (68a) and (68b), we obtain second-order difference equations for the coefficients A_n, B_n, and C_n:

$$\text{(70)}\qquad x\text{-translation:}\quad \begin{aligned} &\widetilde{M}_0(\xi_0)A_1 = 2v_x,\\ &\widetilde{M}_1(\xi_0)A_2 + 2\mathrm{Q}_1(\cosh\xi_0)\,B_1 = 0,\\ &\widetilde{K}_1(\xi_0)A_2 - \frac{1}{2}\mathrm{Q}_1^{(1)}(\cosh\xi_0)\,B_1 = 0, \end{aligned}$$

$$\text{(71)}\qquad y\text{-rotation:}\quad \begin{aligned} &\widetilde{M}_0(\xi_0)A_1 = 0,\\ &\widetilde{M}_1(\xi_0)A_2 + 2\mathrm{Q}_1(\cosh\xi_0)\,B_1 = 2\varpi_y\, c\cosh\xi_0,\\ &\widetilde{K}_1(\xi_0)A_2 - \frac{1}{2}\mathrm{Q}_1^{(1)}(\cosh\xi_0)\,B_1 = \varpi_y\, c\sinh\xi_0, \end{aligned}$$

and for both problems of the x-translation and y-rotation for $n \geq 2$:

$$\widetilde{L}_n(\xi_0)A_{n-1} + \widetilde{M}_n(\xi_0)A_{n+1} + n(n+1)\mathrm{Q}_n(\cosh\xi_0)\, B_n = 0, \tag{72a}$$

$$\begin{aligned} -\frac{c\,\sinh\xi_0}{2(2n-1)}\mathrm{Q}^{(1)}_{n-1}(\cosh\xi_0)A_{n-1} + \frac{c\,\sinh\xi_0}{2(2n+3)}\mathrm{Q}^{(1)}_{n+1}(\cosh\xi_0)A_{n+1}& \\ + \mathrm{Q}^{(2)}_n(\cosh\xi_0)\, C_n = 0&, \end{aligned} \tag{72b}$$

$$\widetilde{N}_n(\xi_0)A_{n-1} + \widetilde{K}_n(\xi_0)A_{n+1} + \frac{1}{2}\mathrm{Q}^{(1)}_n(\cosh\xi_0)\,(-B_n + (n-1)(n+2)\,C_n) = 0, \tag{72c}$$

where the functions $\widetilde{L}_n(\xi)$, $\widetilde{M}_n(\xi)$, $\widetilde{N}_n(\xi)$, and $\widetilde{K}_n(\xi)$ are determined by

$$\begin{aligned} \widetilde{L}_n(\xi_0) &= c\,\frac{n(n-1)}{2n-1}\left(\frac{3}{2}\sinh\xi_0\,\mathrm{Q}^{(1)}_{n-1}(\cosh\xi_0) + n\cosh\xi_0\,\mathrm{Q}_{n-1}(\cosh\xi_0)\right), \\ \widetilde{M}_n(\xi_0) &= c\,\frac{(n+1)(n+2)}{2n+3}\left(-\frac{3}{2}\sinh\xi_0\,\mathrm{Q}^{(1)}_{n+1}(\cosh\xi_0)\right. \\ &\qquad\qquad \left. + (n+1)\cosh\xi_0\,\mathrm{Q}_{n+1}(\cosh\xi_0)\right), \\ \widetilde{N}_n(\xi_0) &= -c\,\frac{(n-1)}{2n-1}\left(\frac{1}{2}\cosh\xi_0\,\mathrm{Q}^{(1)}_{n-1}(\cosh\xi_0) + n\sinh\xi_0\,\mathrm{Q}_{n-1}(\cosh\xi_0)\right), \\ \widetilde{K}_n(\xi_0) &= c\,\frac{(n+2)}{2n+3}\left(-\frac{1}{2}\cosh\xi_0\,\mathrm{Q}^{(1)}_{n+1}(\cosh\xi_0) + (n+1)\sinh\xi_0\,\mathrm{Q}_{n+1}(\cosh\xi_0)\right). \end{aligned}$$

The combination

$$\begin{aligned} (72\text{c})\cdot\mathrm{Q}_n(\cosh\xi_0)\mathrm{Q}^{(2)}_n(\cosh\xi_0) - (72\text{b})\cdot\frac{1}{2}(n-1)(n+2)\mathrm{Q}_n(\cosh\xi_0)\mathrm{Q}^{(1)}_n(\cosh\xi_0)& \\ + (72\text{a})\cdot\frac{1}{2n(n+1)}\mathrm{Q}^{(1)}_n(\cosh\xi_0)\mathrm{Q}^{(2)}_n(\cosh\xi_0)& \end{aligned}$$

reduces to

$$c\left(-\widetilde{A}_{n-1} + \widetilde{A}_{n+1}\right)\widetilde{\delta}_n = 0, \qquad n \geq 2, \tag{73}$$

where $\widetilde{A}_n = \frac{n(n+1)}{2(2n+1)}\,A_n$ and

$$\begin{aligned} \widetilde{\delta}_n = \frac{n(n+1)}{2}&\sinh[2\xi_0]\,(\mathrm{Q}_n(\cosh\xi_0))^3 + \frac{\cosh^2\xi_0}{n(n+1)}\left(\mathrm{Q}^{(1)}_n(\cosh\xi_0)\right)^3 \\ &- \frac{\left(4\coth\xi_0 + (n^2+n-1)\sinh[2\xi_0]\right)}{2n(n+1)}\,\mathrm{Q}_n(\cosh\xi_0)\left(\mathrm{Q}^{(1)}_n(\cosh\xi_0)\right)^2 \\ &- \cosh[2\xi_0]\,(\mathrm{Q}_n(\cosh\xi_0))^2\,\mathrm{Q}^{(1)}_n(\cosh\xi_0). \end{aligned}$$

Since $\widetilde{\delta}_n \neq 0$ for $n \geq 2$, it follows from (73) and (72a)–(72c) that for both x-translation and y-rotation,

$$\widetilde{A}_{n+1} = \widetilde{A}_{n-1}, \qquad n \geq 2, \tag{74a}$$

$$B_n = c\left(\cosh^2\xi_0 - 3\right)\frac{(2n+1)}{n(n+1)}\,\widetilde{A}_{n-1}, \qquad n \geq 2, \tag{74b}$$

$$C_n = -c\,\sinh^2\xi_0\,\frac{(2n+1)}{n(n-1)(n+1)(n+2)}\,\widetilde{A}_{n-1}, \qquad n \geq 2. \tag{74c}$$

Solving (70) and (71), we obtain

$$\text{(75a)}\qquad x\text{-translation:}\qquad \widetilde{A}_1 = \frac{q_1}{2c}, \qquad \widetilde{A}_2 = 0, \qquad B_1 = 0,$$

$$\text{(75b)}\qquad y\text{-rotation:}\qquad \widetilde{A}_1 = 0, \qquad \widetilde{A}_2 = \frac{2q_2}{c}, \qquad B_1 = 2q_2\left(\cosh[2\xi_0] - 2\right).$$

Finally, substituting (74a)–(74c) along with (75a) and (75b) into (69), and using the representations (80) in [32] and also

$$\frac{\sin^2\eta}{\sinh^2\xi\,(\cosh\xi - \cos\eta)} = \sum_{n=2}^{\infty} \frac{(2n+1)}{(n-1)n(n+1)(n+2)}\,\mathrm{Q}_n^{(2)}(\cosh\xi)\,\mathrm{P}_n^{(2)}(\cos\eta),$$

we obtain (66) and (67). The resisting force and torque follow from (48b) and (54b), respectively.

Example 3 (asymmetric translation and rotation of a solid oblate spheroid). Let (ξ, η, φ) be the oblate spheroidal coordinates related to the cylindrical coordinates by

$$\text{(76)}\qquad r = c\cosh\xi\,\sin\eta, \qquad z = c\sinh\xi\,\cos\eta, \qquad \xi \in [0,\infty), \quad \eta \in [0,\pi],$$

where the angular coordinate $\varphi \in [0, 2\pi)$ coincides with the one in (r, φ, z) and c is a metric parameter. In (ξ, η, φ), a solid oblate spheroid with the z-axis of revolution is determined by fixing the coordinate ξ, i.e., $\xi = \xi_0$. For the x-translation (28) and y-rotation (29), let the velocity field be represented by (30a)–(30b).

(i) For the x-translation, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are determined in the region $\xi \geq \xi_0$ by

(77)

$$G_1^{(0)}(\xi,\eta) = -\frac{q_1}{c}\,\frac{(\cos\eta + i\,\tanh\xi\sin\eta)}{\sinh^2\xi + \cos^2\eta},$$

$$G_2^{(0)}(\xi,\eta) = \frac{q_1}{2}\left(\sinh^2\xi_0 + 3\right)\left(\frac{\sinh[2\xi] - i\,\sin[2\eta]}{2\cosh\xi\left(\sinh^2\xi + \cos^2\eta\right)} - \operatorname{arccot}[\sinh\xi]\right),$$

$$G_3^{(1)}(\xi,\eta) = \frac{q_1}{2}\,\cosh^2\xi_0\,\frac{\sin\eta}{\cosh\xi}\,\frac{(\cos\eta + i\,\tanh\xi\sin\eta)}{\sinh^2\xi + \cos^2\eta},$$

where $q_1 = 4v_x\big/\left(\sinh\xi_0 - \left(\sinh^2\xi_0 + 3\right)\operatorname{arccot}[\sinh\xi_0]\right)$. The drag force $F_x = 4\pi\mu c q_1$ follows from (48b) and (77).

(ii) For the y-rotation, the functions $G_1^{(0)}$, $G_2^{(0)}$, and $G_3^{(1)}$ are determined in the region $\xi \geq \xi_0$ by

(78)

$$G_1^{(0)}(\xi,\eta) = \frac{4q_2}{c}\left(-\operatorname{arccot}[\sinh\xi] + \frac{\sinh[2\xi] - i\,\sin[2\eta]}{2\cosh\xi\left(\sinh^2\xi + \cos^2\eta\right)}\right),$$

$$\begin{aligned} G_2^{(0)}(\xi,\eta) = {} & q_2\left(\sinh^2\xi_0 - 3\right)\left(-2\mathrm{Q}_1(-i\sinh\xi)\cos\eta + i\,\mathrm{Q}_1^{(1)}(-i\sinh\xi)\sin\eta\right) \\ & + 2q_2\left(\sinh^2\xi_0 + 3\right)\frac{(\cos\eta + i\,\tanh\xi\sin\eta)}{\sinh^2\xi + \cos^2\eta}, \end{aligned}$$

$$G_3^{(1)}(\xi,\eta) = q_2\cosh^2\xi_0\sin\eta\left(3\mathrm{Q}_1^{(1)}(-i\sinh\xi) - \frac{\sinh[2\xi] - i\,\sin[2\eta]}{\cosh^2\xi\left(\sinh^2\xi + \cos^2\eta\right)}\right),$$

where $q_2 = \varpi_y c \big/ \big(\sinh\xi_0 - \big(\sinh^2\xi_0 - 1\big)\,\mathrm{arccot}[\sinh\xi_0]\big)$, and

$$\begin{aligned}\mathrm{Q}_1(-i\sinh\xi) &= \sinh\xi\ \mathrm{arccot}(\cosh\xi) - 1,\\ \mathrm{Q}_1^{(1)}(-i\sinh\xi) &= \cosh\xi\ \mathrm{arccot}(\cosh\xi) - \tanh\xi.\end{aligned}$$

The resisting torque $T_y = -16\pi\mu c^2 q_2 \cosh[2\xi_0]/3$ follows from (54b) and (78).

Detail. Obtaining (77) and (78) is similar to how (66) and (67) have been obtained.

Similarly, exact solutions to 3D Stokes flow problems for asymmetric motion of a solid spindle, lens, bispheres, and torus of circular cross-section can be obtained in terms of (30a)–(30b) with integral or series representations for k-harmonically analytic functions for corresponding regions; see [35, 34, 33].

In the next two examples, we solve the integral equations (37) and (38) for the x-translation of solid bispheroids (two separate, equal-size spheroids with the same axis of revolution) and for the y-rotation of a solid torus of elliptical cross-section. Exact solutions to these asymmetric Stokes flow problems are available only for two spheres [27] and a torus of circular cross-section [8]. To determine (39) and (40) for the x-translation (28) and y-rotation (29), we have

x-translation:	$f_1(\zeta) = 2iv_x$,	$f_2(\zeta) = -2v_x$,	$f_3(\zeta) = 0$,
y-rotation:	$f_1(\zeta) = -2\varpi_y\overline{\zeta}$,	$f_2(\zeta) = -2\varpi_y z$,	$f_3(\zeta) = 0$.

Example 4 (x-translation of solid bispheroids). Let the centers of bispheroids lie on the z-axis, which is the axis of revolution, and have coordinates $z = \coth 1$ (upper spheroid) and $z = -\coth 1$ (lower spheroid). Let ℓ_+ for $z \geq 0$ (upper spheroid) be parameterized by $r(t) = \varkappa \sin t/\sinh 1$, $z(t) = \coth 1 - \cos t/\sinh 1$, $t \in [0,\pi]$, where $\varkappa$ is a positive parameter. The case $\varkappa = 1$ corresponds to bispheres. For the x-translation (28), we represent $G_1^{(0)}(\zeta(t))$ and $U_3^{(1)}(\zeta(t))$ on ℓ_+ for $z \geq 0$ (upper spheroid) by

$$\begin{aligned} G_1^{(0)}(\zeta(t)) &= \sum_{k=1}^{n} (a_k + i\,b_k)\cos[(k-1)t], \qquad t \in [0,\pi],\\ U_3^{(1)}(\zeta(t)) &= \sum_{k=1}^{n} c_k \cos[(k-1)t], \qquad t \in [0,\pi], \end{aligned} \tag{79}$$

and determine $G_1^{(0)}(\zeta(t))$ and $U_3^{(1)}(\zeta(t))$ on ℓ_+ for $z \leq 0$ (lower spheroid) using $G_1^{(0)}\left(\overline{\zeta}\right) = -\overline{G_1^{(0)}(\zeta)}$ and $U_3^{(1)}\left(\overline{\zeta}\right) = -U_3^{(1)}(\zeta)$ (see Remark 5). The integral equations (37) and (38) can be solved with respect to the coefficients a_k, b_k, and c_k in (79) either by minimizing the total quadratic error of (37) and (38) with (79) or by the collocation method. In both cases, a_k, b_k, and c_k are found from a system of linear equations. Figures 1 and 2 illustrate profiles of the pressure on the surface of the bispheres ($\varkappa = 1$) and bispheroids for $\varkappa = 0.5$ and 2. Table 1 presents the ratio, d_x, of the drag, exerted on one of the two spheroids and calculated by (48a), to the drag of a single same-size spheroid[9] in the x-translation for $\varkappa = 0.5$, 1, and 2.

[9] This spheroid has the size of one spheroid in the bispheroids.

TABLE 1

The ratio, d_x, of the drag, exerted on one of the two solid spheroids, to the drag of a single same-size solid spheroid in the asymmetric translation along the x-axis for $\varkappa = 0.5$, 1, and 2. The case $\varkappa = 1$ corresponds to the bispheres. The value 0.7996 for $\varkappa = 1$ coincides with the one provided in [27].

$\varkappa$	0.5	1.0	2
d_x	0.8486	0.7996	0.7365

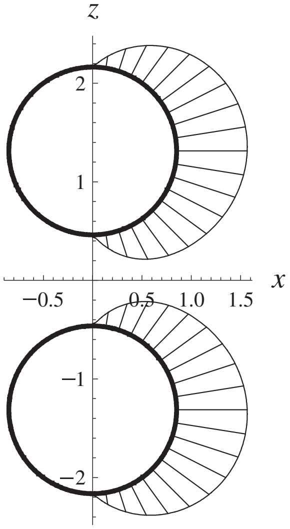
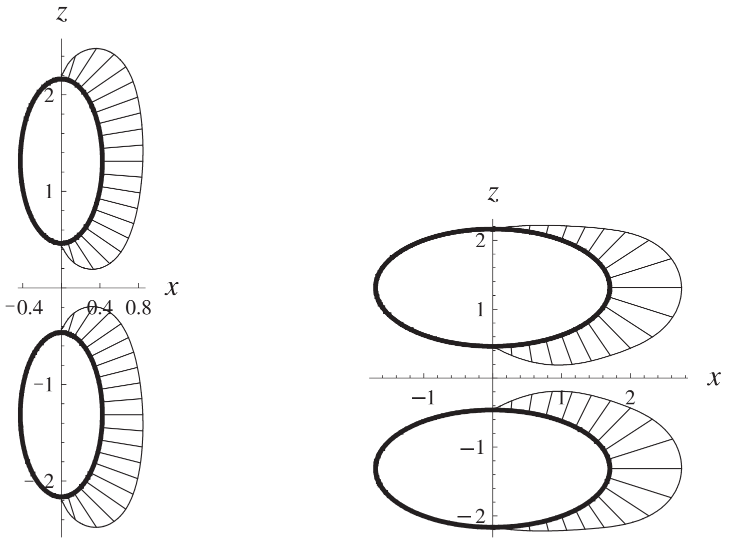


FIG. 1. *Profile of the pressure, $\wp/(2v_x\mu)$ on the surface of the solid bispheres ($\varkappa = 1$) in the asymmetric translation along the x-axis.*



(a) $\wp/(5v_x\mu)$, prolate bispheroids (b) $\wp/(v_x\mu)$, oblate bispheroids

FIG. 2. *Profile of the pressure on the surface of the solid bispheroids in the asymmetric translation along the x-axis:* (a) $\wp/(5v_x\mu)$, *prolate bispheroids* ($\varkappa = 0.5$)*;* (b) $\wp/(v_x\mu)$, *oblate bispheroids* ($\varkappa = 2$).

Table 2

The normalized resisting torque, $t_x = T_x/(216\pi\mu\varpi_y)$, for the solid torus of elliptical cross-section in the asymmetric rotation around the y-axis for $\varkappa = 0.5$, 1, and 2. The case $\varkappa = 1$ corresponds to the torus of circular cross-section. The value 0.5590 *for $\varkappa = 1$ coincides with the one reported in* [8].

$\varkappa$	0.5	1.0	2
t_x	0.4636	0.5590	0.8903

Example 5 (y-rotation of a solid torus of elliptical cross-section). For torus of elliptical cross-section, let ℓ_+ be parameterized by $r(t) = 2 + \cos t$, $z(t) = \varkappa \sin t$, $t \in [-\pi, \pi]$, where $\varkappa$ is a positive parameter. The case $\varkappa = 1$ corresponds to the torus of circular cross-section. For the y-rotation (29), we represent $G_1^{(0)}(\zeta(t))$ and $U_3^{(1)}(\zeta(t))$ on ℓ_+ by

$$\begin{aligned} G_1^{(0)}(\zeta(t)) &= \sum_{k=1}^{n} \left(a_k \cos[(n-1)t] + i\, b_k \sin[k\, t] \right), \qquad t \in [-\pi, \pi], \\ U_3^{(1)}(\zeta(t)) &= \sum_{k=1}^{n} c_k \cos[(k-1)t], \qquad t \in [-\pi, \pi], \end{aligned} \tag{80}$$

for which $G_1^{(0)}\left(\overline{\zeta}\right) = \overline{G_1^{(0)}(\zeta)}$ and $U_3^{(1)}\left(\overline{\zeta}\right) = U_3^{(1)}(\zeta)$ on ℓ_+ (see Remark 5). The integral equations (37) and (38) can be solved with respect to the coefficients a_k, b_k, and c_k by the same methods as in Example 4. Figures 3, 4, and 5 illustrate profiles of the pressure on the surface of the torus of elliptical cross-section for $\varkappa = 1$, 0.5, and 2, respectively. Table 2 presents the ratio of the resisting torque, exerted on the torus and calculated by (54a), to the resisting torque, $216\pi\mu\varpi_y$, of the circumscribed sphere of radius 3 in the y-rotation for $\varkappa = 0.5$, 1, and 2.

Appendix. Auxiliary results for *k*-harmonically analytic functions. This section presents two auxiliary results dealing with implementation of the necessary and sufficient condition for a function $G^{(k)}$ to be k-harmonically analytic in an outer region and vanishing at infinity.

Let ℓ_+, ℓ, $\mathcal{D}^+$, $\mathcal{D}_0^+$, $\mathcal{D}^-$, and $\mathcal{D}_0^-$ be defined as in section 2, where $\mathcal{D}_0^+$ consists of disjoint simply connected subregions $\mathcal{D}_j^+$, $1 \le j \le m$. Also let $\widehat{\ell}_j$ be the part of the boundary ℓ_+ that corresponds to $\mathcal{D}_j^+$. It is either a closed curve or open curve with the endpoints lying on the z-axis. Obviously, $\ell_+ = \bigcup_{j=1}^m \widehat{\ell}_j$. We first establish the following result.

Proposition 9. *Let $G^{(k)}(\zeta) = U^{(k)}(\zeta) + i\,V^{(k+1)}(\zeta)$ be a k-harmonically analytic function in $\mathcal{D}_j^+$, $1 \le j \le m$, and let the functions $\left(\frac{\partial}{\partial r} - \frac{k}{r}\right) U^{(k)}$, $\frac{\partial}{\partial z} U^{(k)}$, $\left(\frac{\partial}{\partial r} + \frac{k+1}{r}\right) V^{(k+1)}$, and $\frac{\partial}{\partial z} V^{(k+1)}$ be continuously differentiable in $\mathcal{D}_j^+$, $1 \le j \le m$.*

(i) *If $U^{(k)}\big|_{\ell_+} = 0$, then $V^{(k+1)}(\zeta) = a_j\, r^{-k-1}$ in $\mathcal{D}_j^+$, where a_j is a real-valued constant.*

(ii) *If $V^{(k+1)}\big|_{\ell_+} = 0$, then $U^{(k)}(\zeta) = b_j\, r^k$ in $\mathcal{D}_j^+$, where b_j is a real-valued constant.*

Proof. First we prove (i). If $\widehat{\ell}_j$ is an open curve, we close it by connecting its

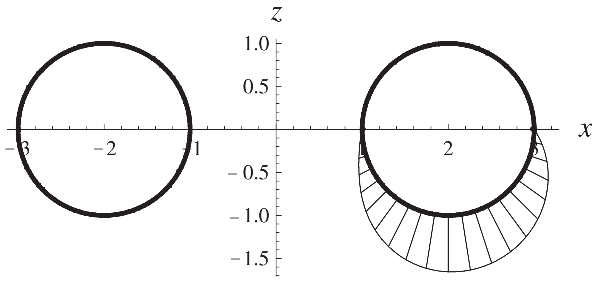


FIG. 3. *Profile of the pressure, $\wp/(4\varpi_y\mu)$, on the surface of the solid torus of circular cross-section ($\varkappa = 1$) in the asymmetric rotation around the y-axis (for $z > 0$, the profile is antisymmetric).*

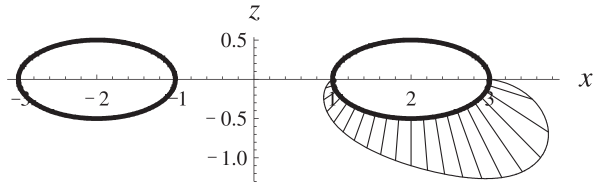


FIG. 4. *Profile of the pressure, $\wp/(4\varpi_y\mu)$, on the surface of the solid torus of elliptical cross-section with $\varkappa = 0.5$ in the asymmetric rotation around the y-axis (for $z > 0$, the profile is antisymmetric).*

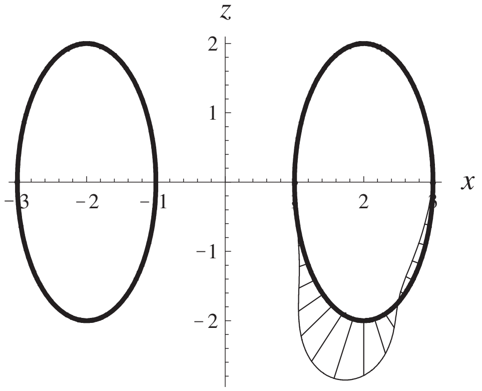


FIG. 5. *Profile of the pressure, $\wp/(4\varpi_y\mu)$, on the surface of the solid torus of elliptical cross-section with $\varkappa = 2$ in the asymmetric rotation around the y-axis (for $z > 0$, the profile is antisymmetric).*

endpoints by the segment on the z-axis. By Green's theorem, we have

$$\oint_{\widehat{\ell}_j} r\,U^{(k)}\left(\left[\left(\frac{\partial}{\partial r}-\frac{k}{r}\right)U^{(k)}\right]dz-\left[\frac{\partial}{\partial z}U^{(k)}\right]dr\right)$$

$$=\iint_{\mathcal{D}_j^+} r\,U^{(k)}\,\Delta_k U^{(k)}\,drdz+\iint_{\mathcal{D}_j^+} r\left(\left[\frac{\partial}{\partial z}U^{(k)}\right]^2+\left[\left(\frac{\partial}{\partial r}-\frac{k}{r}\right)U^{(k)}\right]^2\right)drdz.$$

Since $U^{(k)}\big|_{\ell_+}=0$ implies $U^{(k)}\big|_{\widehat{\ell}_j}=0$, the left-hand side of this equation is zero. Also the first integral in the right-hand side vanishes, since $U^{(k)}$ is a k-harmonic function, i.e., $\Delta_k U^{(k)} = 0$. Thus, $\left[\frac{\partial}{\partial z} U^{(k)}\right]^2 + \left[\left(\frac{\partial}{\partial r} - \frac{k}{r}\right) U^{(k)}\right]^2 \equiv 0$ in $\mathcal{D}_j^+$, whence $U^{(k)} = c_j\, r^k$, where c_j is a real-valued constant. However, from $U^{(k)}\big|_{\widehat{\ell}_j} = 0$, it follows that $c_j = 0$, and thus, $U^{(k)} \equiv 0$ in $\mathcal{D}_j^+$. Then, from (12), we have $V^{(k+1)} = a_j\, r^{-k-1}$ in $\mathcal{D}_j^+$, where a_j is a real-valued constant, and statement (i) is proved.

Statement (ii) is proved similarly. Observing that

$$\oint_{\widehat{\ell}_j} r\, V^{(k+1)} \left(\left[\left(\frac{\partial}{\partial r} + \frac{k+1}{r} \right) V^{(k+1)} \right] dz - \left[\frac{\partial}{\partial z} V^{(k+1)} \right] dr \right)$$
$$= \iint_{\mathcal{D}_j^+} r\, V^{(k+1)}\, \Delta_{k+1} V^{(k+1)}\, dr dz$$
$$+ \iint_{\mathcal{D}_j^+} r \left(\left[\frac{\partial}{\partial z} V^{(k+1)} \right]^2 + \left[\left(\frac{\partial}{\partial r} + \frac{k+1}{r} \right) V^{(k+1)} \right]^2 \right) dr dz,$$

and arguing as in the proof of part (i), we conclude that $V^{(k+1)} \equiv 0$ in $\mathcal{D}_j^+$, and the result follows. □

The necessary and sufficient condition for a function $G^{(k)}(\zeta) = U^{(k)}(\zeta) + i\, V^{(k+1)}(\zeta)$ to be k-harmonically analytic in the outer region $\mathcal{D}^-$ and vanishing at infinity follows from the generalized Sokhotski–Plemelj formulae and can be written in the form

$$G^{(k)}(\zeta) + \frac{1}{\pi i} \oint_{\ell} G^{(k)}(\tau) \mathcal{W}^{(k)}(\zeta, \tau)\, d\tau = 0, \quad \zeta \in \ell, \tag{81}$$

where ℓ is the boundary of $\mathcal{D}^-$.

Proposition 10. *Let $\mathcal{D}^-$ be symmetric with respect to the r-axis.*

(i) *If $G^{(k)}\left(\overline{\zeta}\right) = \overline{G^{(k)}(\zeta)}$, then* (81) *is equivalent to*

$$U^{(k)}(\zeta) + \operatorname{Re}\left[\frac{1}{\pi i} \oint_{\ell} G^{(k)}(\tau) \mathcal{W}^{(k)}(\zeta, \tau)\, d\tau \right] = 0, \quad \zeta \in \ell. \tag{82}$$

(ii) *If $G^{(k)}\left(\overline{\zeta}\right) = -\overline{G^{(k)}(\zeta)}$, then* (81) *is equivalent to*

$$V^{(k+1)}(\zeta) + \operatorname{Im}\left[\frac{1}{\pi i} \oint_{\ell} G^{(k)}(\tau) \mathcal{W}^{(k)}(\zeta, \tau)\, d\tau \right] = 0, \quad \zeta \in \ell. \tag{83}$$

Proof. The integral equations (82) and (83) are the real and imaginary parts of (81), respectively, and thus, any solution to (81) solves (82) and (83). We need to show that under the given conditions, the converse is also true.

First we prove statement (i). Let $G_*^{(k)}(\zeta)$, $\zeta \in \ell$, be a solution to (82) such that $G_*^{(k)}\left(\overline{\zeta}\right) = \overline{G_*^{(k)}(\zeta)}$, and let a k-harmonically analytic function $\Psi^+(\zeta)$ in $\operatorname{int} \mathcal{D}^+$ (the complement of $\mathcal{D}^-$) be determined by the generalized Cauchy-type integral

$$\Psi^+(\zeta) = \frac{1}{2\pi i} \oint_{\ell} G_*^{(k)}(\tau) \mathcal{W}^{(k)}(\zeta, \tau)\, d\tau, \quad \zeta \in \operatorname{int} \mathcal{D}^+.$$

In the case when ζ approaches ℓ from within the inner region $\mathcal{D}^+$, the boundary value of Ψ^+ on ℓ is determined by the corresponding generalized Sokhotski–Plemelj formula

$$\Psi^+(\zeta) = \frac{1}{2} G_*^{(k)}(\zeta) + \frac{1}{2\pi i} \oint_{\ell} G_*^{(k)}(\tau) \mathcal{W}^{(k)}(\zeta, \tau)\, d\tau, \quad \zeta \in \ell.$$

Since $G_*^{(k)}$ is a solution to (82), $\operatorname{Re}\Psi^+(\zeta) = 0$ on ℓ, whence $\operatorname{Re}\Psi^+(\zeta) = 0$ on ℓ_+. Then, by Proposition 9(i), we have $\operatorname{Im}\Psi^+(\zeta) = a_j\, r^{-k-1}$ in $\mathcal{D}_j^+$, where a_j is a real-valued constant, and $\mathcal{D}_j^+$, $1 \le j \le m$, are disjoint simply connected subregions of $\mathcal{D}_0^+$ such that $\mathcal{D}_0^+ = \bigcup_{j=1}^m \mathcal{D}_j^+$. However, since $G_*^{(k)}\left(\overline{\zeta}\right) = \overline{G_*^{(k)}(\zeta)}$, we have $\operatorname{Im}\Psi^+\left(\overline{\zeta}\right) = -\operatorname{Im}\Psi^+(\zeta)$, and consequently, $a_j = 0$ and $\Psi^+(\zeta) \equiv 0$ on ℓ_+. Thus, $G_*^{(k)}$ satisfies (81), and under the given condition the equivalence of (82) and (81) follows.

Statement (ii) is proved similarly. In this case, we have the same function $\Psi^+(\zeta)$ and use Proposition 9(ii). □

The necessary and sufficient condition for a function $f(\zeta)$ to be (ordinary) analytic in the outer region $\mathcal{D}^-$ and vanishing at infinity is given by $f(\zeta) + \frac{1}{\pi i}\oint_\ell \frac{f(\tau)}{\tau-\zeta}\,d\tau = 0$, $\zeta \in \ell$, which reduces to the nonsingular integral equation $2f(\zeta) + \frac{1}{\pi i}\oint_\ell \frac{f(\tau)-f(\zeta)}{\tau-\zeta}\,d\tau = 0$, $\zeta \in \ell$, with the fact that $c - \frac{1}{\pi i}\oint_\ell \frac{c}{\tau-\zeta}\,d\tau = 0$, $\zeta \in \ell$, for an arbitrary constant c. The next proposition extends this well-known technique for k-harmonically analytic functions, where the function r^k assumes the role of constant.

PROPOSITION 11. *The singular integral equation* (81) *reduces to*

$$\begin{aligned} 2\, r^k\, G^{(k)}(\zeta) + \frac{1}{\pi i}\int_{\ell_+} &\Big(r^k\, G^{(k)}(\tau) - r_1^k\, G^{(k)}(\zeta)\Big)\frac{\Omega_+^{(k)}(\zeta,\tau)}{\tau-\zeta}\,d\tau \\ &- \Big(r^k\,\overline{G^{(k)}(\tau)} - r_1^k\, G^{(k)}(\zeta)\Big)\frac{\Omega_-^{(k)}(\zeta,\tau)}{\overline{\tau}+\zeta}\,d\overline{\tau} = 0, \qquad \zeta \in \ell_+, \end{aligned} \tag{84}$$

where $r_1 = \operatorname{Re}\tau$. *Equation* (84) *has only logarithmic singularity because of the function* $\Omega_-^{(k)}$*; see Remark* 2 *in* [32].

Proof. Since r^k is a k-harmonically analytic function in $\mathcal{D}^+$ (the symmetry condition (15) holds), the generalized Sokhotski–Plemelj formula, corresponding to the case when ζ approaches ℓ from within $\mathcal{D}^+$, implies that

$$r^k - \frac{1}{\pi i}\oint_\ell r_1^k\, \mathcal{W}^{(k)}(\zeta,\tau)\,d\tau = 0, \qquad \zeta \in \ell. \tag{85}$$

The combination $r^k \cdot (81) + G^{(k)}(\zeta)\cdot(85)$ reduces to (84). □

Acknowledgment. We are grateful to the anonymous referees for their valuable comments and suggestions, which helped to improve the quality of the paper.

REFERENCES

[1] A. Ya. Alexandrov and Yu. I. Soloviev, *Three-Dimensional Problems of the Theory of Elasticity*, Nauka, Moscow, 1978 (in Russian).

[2] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, UK, 2000.

[3] L. Bers, *Theory of Pseudo-Analytic Functions*, Institute for Mathematics and Mechanics, New York University, New York, 1953.

[4] L. Bers, *An outline of the theory of pseudoanalytic functions*, Bull. Amer. Math. Soc., 62 (1956), pp. 291–331.

[5] W. D. Collins, *A note on the axisymmetric Stokes flow of viscous fluid past a spherical cap*, Mathematika, 10 (1963), pp. 72–78.

[6] W. R. Dean and M. E. O'Neill, *A slow motion of viscous fluid caused by the rotation of a solid sphere*, Mathematika, 10 (1963), pp. 13–24.

[7] F. D. Gakhov, *Boundary Value Problems*, Pergamon Press, Oxford, New York, 1966.

[8] S. L. Goren and M. E. O'Neill, *Asymmetric creeping motion of an open torus*, J. Fluid Mech., 101 (1980), pp. 97–110.

[9] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics*, Springer, New York, 1983.

[10] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1945.

[11] N. I. Muskhelishvili, ed., *Some Basic Problems of the Mathematical Theory of Elasticity*, Springer, New York, 1977.

[12] N. I. Muskhelishvili, *Singular Integral Equations: Boundary Problems of Function Theory and Their Applications to Mathematical Physics*, 2nd ed., Dover, New York, 1992.

[13] A. Nir and A. Acrivos, *On the creeping motion of two arbitrary-sized touching spheres in a linear shear field*, J. Fluid Mech., 59 (1973), pp. 209–223.

[14] H. Oberbeck, *Über stationäre flüssigkeitsbewegungen mit berücksichtigung der innere reibung*, J. Reine Angew. Math., 81 (1876), pp. 62–80.

[15] M. E. O'Neill, *Exact solutions of the equations of slow viscous flow generated by the asymmetrical motion of two equal spheres*, Appl. Sci. Res., 21 (1969), pp. 452–466.

[16] L. E. Payne and W. H. Pell, *The Stokes flow problem for a class of axially symmetric bodies*, J. Fluid Mech., 7 (1960), pp. 529–549.

[17] W. H. Pell and L. E. Payne, *On Stokes flow about a torus*, Mathematika, 7 (1960), pp. 78–92.

[18] W. H. Pell and L. E. Payne, *The Stokes flow about a spindle*, Quart. Appl. Math., 18 (1960/1961), pp. 257–262.

[19] G. N. Položii, *Theory and Application of p-Analytic and (p, q)-Analytic Functions*, 2nd ed., Naukova Dumka, Kiev, 1973 (in Russian).

[20] C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge University Press, Cambridge, UK, 1992.

[21] C. Pozrikidis, *Introduction to Theoretical and Computational Fluid Dynamics*, The Clarendon Press, Oxford University Press, New York, 1997.

[22] M. Stimson and G. B. Jeffery, *The motion of two-spheres in a viscous fluids*, Proc. Roy. Soc. Lond. Ser. A, 111 (1926), pp. 110–116.

[23] G. G. Stokes, *On the effect of the internal friction of fluids on the motion of pendulums*, Trans. Cambridge Philos. Soc., 9 (1850), pp. 8–106.

[24] H. Takagi, *Slow viscous flow due to the motion of a closed torus*, J. Phys. Soc. Japan, 35 (1973), pp. 1225–1227.

[25] A. F. Ulitko, *Vectorial Decompositions in the Three-Dimensional Theory of Elasticity*, Akademperiodika, Kiev, 2002 (in Russian).

[26] I. N. Vekua, *Generalized Analytic Functions*, Pergamon Press, London, Addison–Wesley, Reading, MA, 1962.

[27] S. Wakiya, *Slow motion of a viscous fluid around two spheres*, J. Phys. Soc. Japan, 22 (1967), pp. 1101–1109.

[28] S. Wakiya, *On the exact solution of the Stokes equations for a torus*, J. Phys. Soc. Japan, 37 (1974), pp. 780–783.

[29] G. K. Youngren and A. Acrivos, *Stokes flow past a particle of arbitrary shape: A numerical method of solution*, J. Fluid Mech., 69 (1975), pp. 377–403.

[30] M. Zabarankin, *Asymmetric three-dimensional Stokes flows about two fused equal spheres*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 463 (2007), pp. 2329–2349.

[31] M. Zabarankin, *Asymmetric creeping motion of a rigid spindle-shaped body in a viscous fluid*, SIAM J. Appl. Math., 68 (2007), pp. 461–485.

[32] M. Zabarankin, *The framework of k-harmonically analytic functions for three-dimensional Stokes flow problems, Part* I, SIAM J. Appl. Math., 69 (2008), pp. 845–880.

[33] M. Zabarankin and P. Krokhmal, *Generalized analytic functions in* 3D *Stokes flows*, Quart. J. Mech. Appl. Math., 60 (2007), pp. 99–123.

[34] M. Zabarankin and A. F. Ulitko, *Hilbert formulas for r-analytic functions and Stokes flow about a biconvex lens*, Quart. Appl. Math., 64 (2006), pp. 663–693.

[35] M. Zabarankin and A. F. Ulitko, *Hilbert formulas for r-analytic functions in the domain exterior to spindle*, SIAM J. Appl. Math., 66 (2006), pp. 1270–1300.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 909–932

ESTIMATING A GREEN'S FUNCTION FROM "FIELD-FIELD" CORRELATIONS IN A RANDOM MEDIUM[*]

MAARTEN V. DE HOOP[†] AND KNUT SOLNA[‡]

Abstract. Traditional imaging methods use coherent signals as data. Here, we discuss recent developments in imaging that aim at exploiting as data incoherent noisy signals that are not associated with well-defined arrival times. Indeed, signal constituents that in a classical setting may be regarded as noise may contain important information about the medium to be imaged. We show how it is possible to use the statistics of such noisy signals, specifically, the second-order statistics, for imaging. We consider two particular situations: first, the estimation of an ("empirical") Green's function from noisy signals which can subsequently be used in imaging; second, the localization of a cluster of random sources from noisy signals (passive imaging). The analysis presented here is based on assuming a remote sensing scaling and the paraxial approximation, and it uses in part the results set forth in Papanicolaou, Ryzhik, and Solna [*SIAM J. Appl. Math.*, 64 (2004), pp. 1133–1155] that relate to time-reversal, statistical stability, and superresolution. Robustness with respect to modeling assumptions is illustrated by considering other scaling regimes also. We demonstrate how the estimation problem and its robustness can be considered as a dual to that of time-reversal and stable superresolution. We obtain a novel analysis and foundation for the use of ambient seismic noise in body-wave (tomographic) imaging, motivated by the recent successes of surface-wave tomography using ambient seismic noise.

Key words. empirical Green's function, random medium, paraxial approximation, statistical stability

AMS subject classifications. 34F05, 34E10, 37H10, 60H20

DOI. 10.1137/070701790

1. Introduction.

1.1. Time-reversal and cross-correlation–based imaging.

Recent work on time-reversal of waves in a random medium has shown that medium fluctuations are not necessarily detrimental to, but may in fact enhance various operations with, waves. This has been analyzed mathematically as well as demonstrated experimentally [4, 20, 22, 25, 30, 31, 33]. In classical time-reversal, the wave received by an active transducer (receiver-emitter) array is recorded and then re-emitted into the configuration time-reversed; that is, the tails of the recorded signals are sent first. In the absence of absorption, the re-emitted signal will propagate back toward the source and focus, approximately, on it. This phenomenon has a large number of applications, in inverse problems, medical imaging, remote sensing, target identification, and secure communication, for instance. Here, we discuss the notion of "field-field" cross-correlations associated with noise observed at pairwise distinct receivers, to obtain an "empirical" Green's function. This notion is naturally related to the time-reversal mentioned above [13, 14]. We will give a precise characterization of the "empirical" Green's function, which can subsequently be used for imaging the interrogated

[*] Received by the editors September 4, 2007; accepted for publication (in revised form) July 1, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/siap/69-4/70179.html

[†] Center for Computational and Applied Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907 (mdehoop@math.purdue.edu).

[‡] Department of Mathematics, University of California at Irvine, Irvine, CA 92697 (ksolna@math.uci.edu). This author's research was supported in part by DARPA grant N00014-02-1-0603, ONR grant N00014-02-1-0090, NSF grant 0307011, and the Sloan Foundation.

medium. To this end, we consider a configuration consisting of a randomly heterogeneous halfspace, possibly containing a scatterer, which is exposed to random *noisy sources* concentrated at the surface (that is, the top of the halfspace). The idea is to use *cross-correlations* between a set of measurements of noise to infer information about the medium in the halfspace as well as about the location of the scatterer.

With seismology as a key application in mind, the goal is to obtain an image of Earth's interior from all possible signals recorded at an array of receiver stations. Robust imaging requires, essentially, a regular distribution of sources. Effectively using receivers as sources through the mentioned "field-field" cross-correlations, this requirement can be satisfied, even where deterministic sources (earthquakes) are necessarily absent. The idea of using ambient noise for the retrieval of a body-wave reflection response, in a planarly layered medium, dates back to Claerbout [8]. He also conjectured that, in general media, cross-correlating ambient noise traces from two locations recaptures the wavefield at one of the locations, excited by a point source at the other location. An early example of a field application was reported by Scherbaum [37], who analyzed auto-correlations of recordings of low-magnitude earthquakes and generated *pseudoreflection seismograms*. Moreover, the cross-correlation method has been developed in helioseismology [16]. In recent years, the understanding of how cross-correlating diffuse fields recaptures the Green's function has been a topic of research [46, 42]. Cross-correlating (diffuse) coda waves [7] and ambient seismic noise [38, 39] resulted in the retrieval of surface waves observed at one station and excited at the other station; for a detailed study, see Yao, van der Hilst, and deHoop [43]. Furthermore, *turning* body waves have been observed in cross-correlating ambient noise [35]; in an exploration seismology setting, *reflected* body waves have also been recovered by cross-correlations [15]. The exploitation of a scattering medium in capturing the Green's function by field-field cross-correlations was studied by Derode et al. [14]. The mathematical analysis of field-field cross-correlations in the mentioned setting from the point of view of heterogeneous media has just begun. See [26] for analysis of the case with a layered random medium and [3] for a recent analysis based on the semiclassical approximation and that explicitly discusses stability of the traveltime estimates associated with the correlations which requires a careful analysis of the time scales involved. In [10], field-field cross-correlations from ambient noise, which is not necessarily localized or directional, are analyzed in the high-frequency limit, again making use of Egorov's theorem of microlocal analysis. Here, motivated by applications to propagation in the heterogeneous earth and the atmosphere, we will consider the far field regime and consider a stochastic modeling approach. We exploit the random fluctuations in the medium and consider a variety of (scaling) regimes. In the above, noisy sources are passive sources, while the medium is subjected to random fluctuations. Controlled sources can be used to retrieve the Green's function from cross-correlations of wavefields observed at pairwise distinct receivers as well. Time-reversal and Rayleigh's reciprocity relation can be combined to easily identify the cross-correlation as the Green's function in the interior of a compact domain with controlled sources (everywhere) on its boundary without knowledge of the (deterministic) medium [12]; applications of this general result in borehole seismic imaging can be found in [28, 47, 44]. Insights in extending this situation to the case with decorrelating random sources (on the boundary) can be found, for example, in [45].

A proper estimate of the Green's function between a set of locations can be used for robust imaging from ambient noise [27, 39, 43] and for encoding signals also. The technique of cross-correlations is furthermore relevant in the context of

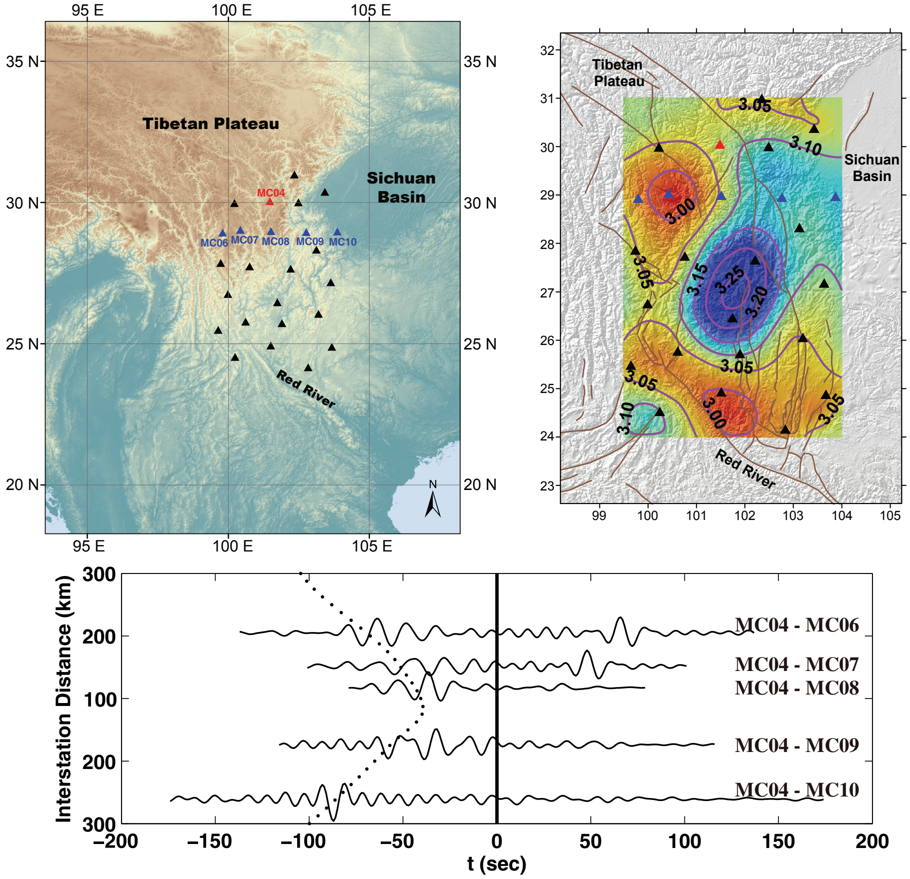


FIG. 1.1. *Top left: a map and the array of receivers. Top right: phase velocities at $T = 10\,s$ (obtained from the Green's function estimates between all pairs of receivers, and given in* [43]*); these phase velocities play the role of c_0. Bottom: Green's function estimates from* 10 *months of data cross-correlations. The positive times correspond to the Green's function from the "source" station MC*04 *(red) to receivers MC∗∗ (blue), while the negative times correspond to the Green's function from the "source" stations MC∗∗ (blue) to receiver station MC*04 *(red). We observe an asymmetry (in time) that will appear in the analysis presented here also. The dotted line indicates the traveltimes computed from the top right figure.*

communication in a waveguide [36] and synchronization of transducer array elements. Detection based on cross-correlations in a noisy environment is discussed in [1] and interferometric imaging in [6]. The theory for analysis of cross-correlations in layered media is presented in [25] with applications presented in, for instance, [23, 24]. While current studies relating to the heterogeneous earth mostly make use of surface-wave contributions to the Green's function estimate, we emphasize, here, the importance of understanding the behavior of body waves for future applications. In Figure 1.1 we illustrate surface-wave contributions to the Green's function estimate over an array in Southeastern Tibet (obtained from [43], upon cross-correlating noise between receiver pairs over a 10 month period).

In this paper, we analyze *estimation based on incoherent waves* in the context of the paraxial approximation and the associated Wigner distribution. We prove that,

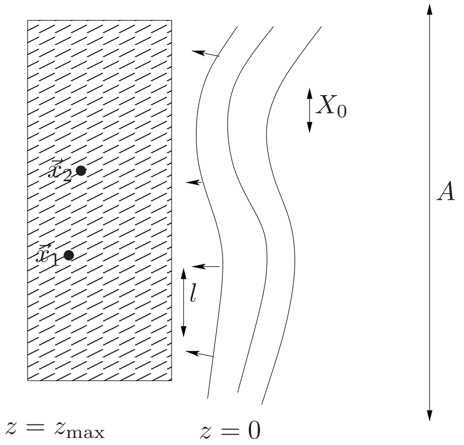


FIG. 1.2. *The experimental setup: A slab of random medium is located in* $(0, z_{\max})$. *In the plane* $z = 0$ *a set of random sources probe the medium, and we record the signal transmitted to the two points of observation* $\vec{x}_1$ *and* $\vec{x}_2$.

in principle, the Green's function can be recovered from cross-correlations, up to a filter that depends only mildly on the medium realization. Moreover, we show that much better estimates (when the Green's function is better resolved) may be obtained in a randomly inhomogeneous medium than in a deterministic (quasi-)homogeneous medium, as a consequence of the wider angular spread in the phase-space representation of a wave in the random medium; this is in agreement with the results of the experiment described in [13]. The enhanced resolution occurs due to an exponential damping factor that appears in the analysis of the cross-correlation and that involves the structure function of the medium; thus, relatively strong disorder gives relatively high resolution. The damping factor was also responsible for superresolution in the time-reversal experiment elucidated in [33], revealing an intrinsic connection. It appears to be important that the signals generating the cross-correlations are subjected to frequency-bandlimitation: The low frequencies must be removed to obtain accurate estimates.

Furthermore, we discuss the localization of a cluster of random sources from noisy signals, a problem which is closely related to the Green's function estimation.

1.2. Configuration and procedure. We consider a configuration and an experiment similar to the one described in [13], while motivated by the procedure and study described in [43]; the configuration is illustrated in Figure 1.2. We use $\mathbf{x} \in \mathbb{R}^d$, with $d \in \{1, 2\}$ denoting the lateral spatial dimension, to represent the "lateral" coordinate(s) and $z \in \mathbb{R}_{\geq 0}$ to represent the "principal" or "depth" coordinate; we write $\vec{x} = (z, \mathbf{x})$. The above-mentioned halfspace contains a heterogeneous slab, the random medium, with a large extent in the lateral directions and supported in $(0, z_{\max})$ in the principal direction. The sources are concentrated in the plane $z = 0$ and are independent of the random medium; they model the ambient noise field. The sources are incoherent and statistically stationary. The aperture, or lateral extent, of the source field is denoted A, and its correlation length X_0, while the correlation length of the random medium is denoted l. We observe the signal, u say, that is due to the noisy sources at the points, $\vec{x}_1$ and $\vec{x}_2$, which may be inside or outside ($z > z_{max}$) the random medium, over the time interval $(0, T)$. The key quantity considered is the cross-correlation function,

$$\mathcal{C}(\tau) = \frac{1}{T}\int_0^T u(t,\vec{x}_1)u(t+\tau,\vec{x}_2)\,dt\,, \tag{1.1}$$

for a large time window $(0,T)$. We will also consider the situation with more than two points of observation: $\vec{x}_1,\ldots,\vec{x}_N$, $N \geq 3$.

1.3. Outline. The outline of the paper is as follows. In section 1.2 we discussed the configurational setup that we will consider. In section 2 we describe the modeling of the sources, the medium, and the stochastic paraxial wave equation formulation. Using this model, we analyze the estimation based on (1.1) in section 3. The main result shows that the cross-correlations give the Green's function blurred by a statistically stable "filter"; cf. (3.19). A striking property of the filter is that its support may be much smaller in a random medium than in a (quasi-)homogeneous medium, which is the counterpart of superresolution in this context. In section 4 we develop an approach to localizing a cluster of random sources from noisy signals. The main aspects of the results we derive are general, and we demonstrate this by discussing different scaling regimes in section 5. In subsection 5.4 we touch upon the scaling regime anticipated in applying the analysis to Southeastern Tibet for the estimation of body-wave constituents from ambient noise. We provide some numerical examples in section 6, and concluding remarks in section 7.

2. High-frequency paraxial regime and modeling.

2.1. The random sources. We shall model the ambient or background far field noise in terms of a random field. The impinging noise will be modeled as an initial condition in the plane $z = 0$ supplementing the paraxial evolution equation to be introduced in the next subsection; see Figure 1.2.

Let ν be a random field in $\mathbb{R}\times\mathbb{R}^d$, and χ be a smooth, deterministic, envelope function. We assume that ν has zero mean, is isotropic in $\mathbf{x}$ and stationary, and is independent of the medium with spectrum,

$$\mathbb{E}[\nu(t',\mathbf{x}')\nu(t'+\tau,\mathbf{x}'+\Delta\mathbf{x})] = C_0\left(\tau,\Delta\mathbf{x}\right), \tag{2.1}$$

and with rapidly decaying correlations. The noisy sources are then collectively modeled as

$$Y(t,\mathbf{x}) = \sigma_y \nu\left(\frac{t}{T_0},\frac{\mathbf{x}}{X_0}\right)\chi\left(\frac{\mathbf{x}-\mathbf{x}_c}{A}\right). \tag{2.2}$$

The envelope function χ models the locality of the sources. We shall consider both an extended source field, with χ having a "large" support, and a concentrated source field, with χ having a "small" support. It is convenient to introduce a characteristic wavelength scale, λ_0, associated with the noise source spectrum,

$$\lambda_0 = c_0 T_0 = \frac{2\pi c_0}{\omega_0} = \frac{2\pi}{k_0}\,, \tag{2.3}$$

where T_0 is a characteristic time scale associated with the temporal noise correlations (cf. (2.2)) and c_0 background or homogenized medium wavespeed. Note that the central wavenumber is defined by

$$k_0 = \frac{\omega_0}{c_0}\,.$$

2.2. The parabolic wave equation. In recent decades the parabolic or paraxial wave equation has emerged as the primary tool to describe small scale scattering situations as they appear in radiowave propagation, radar, remote sensing, propagation in urban environments, and in underwater acoustics [29, 32, 41], as well as in propagation problems in the earth's crust [9]. The paraxial equation models wave propagation if the dominant scattering occurs in the direction(s) transverse to a principal propagation direction. Here, we take this model as our starting point and consider propagation in a random medium in the regime of waves propagating over distances that are large compared to the correlation length of the random inhomogeneities and the characteristic wavelength. The relevant wavelength scale is determined by the support of the noise auto-covariance function.

Fundamental to the problem at hand is the role of *scales*. Different scaling relations will give rise to different qualitative behavior of the estimation of the Green's function. The important scales are the following:

- z_1, the characteristic depth (longitudinal distance) from noisy sources to the recordings;
- A, the characteristic size for the support of noisy sources collectively;
- T_0, X_0, the temporal and lateral (spatial) extent of the noise spectrum;
- σ_c, the relative magnitude of medium fluctuations; see (2.5);
- l, the correlation length of the (isotropic) medium fluctuations; see (2.5).

The correlation length corresponds to the dominant spatial scale at which the medium fluctuates, and it typically defines the microscale in the problem.

To introduce the paraxial wave approximation, we consider first the wave equation governing the propagation of acoustic waves:

$$\frac{1}{c^2(\vec{x})}\frac{\partial^2 u}{\partial t^2} - \triangle u = 0\,, \quad t \in \mathbb{R}\,, \quad \vec{x} \in \mathbb{R}^{d+1}\,. \tag{2.4}$$

Here, the slowness squared, $c^{-2}(z, \mathbf{x})$, is given by

$$c^{-2}(z, \mathbf{x}) = c_0^{-2}\left[1 + \sigma_c\,\mu\left(\frac{z}{l}, \frac{\mathbf{x}}{l}\right)\right], \tag{2.5}$$

in which μ is a random field modeling the medium fluctuations; c_0 denotes the (deterministic) background wavespeed. In the regime of homogenization, with relatively rapidly fluctuating medium variations, the effective wavespeed is c_0. The regime of homogenization corresponds to the case when the wavelength is large relative to the correlation length of the medium fluctuations, and the propagation distance is on the order of the wavelength. However, in a regime of large propagation distances the effect of the randomness will build up, and this phenomenon will be captured by a random potential, namely through μ, in the paraxial wave equation. We shall here assume that the background wavespeed is constant; see [40] for a discussion of the case with a variable background.

Because "locally" the waves sense a homogeneous medium, it is common practice to introduce the following Fourier transform incorporating the centering in a frame moving with the effective wavespeed,

$$u(t, z, \mathbf{x}) = \frac{1}{2\pi}\int e^{\mathrm{i}\omega(z/c_0 - t)}\psi(z, \mathbf{x}, \omega/c_0)\,d\omega\,, \tag{2.6}$$

so that the complex amplitude $\psi(z, \mathbf{x}, k)$ satisfies the Helmholtz equation

$$2\mathrm{i}k\frac{\partial\psi}{\partial z} + \triangle_{\mathbf{x}}\psi + k^2(n^2 - 1)\psi = -\frac{\partial^2\psi}{\partial z^2}\,, \tag{2.7}$$

with $k = \omega/c_0$ being the wavenumber and $n = n(z, \mathbf{x}) = c_0/c(z, \mathbf{x})$ the random index of refraction relative to the background wavespeed c_0. The fluctuations in the refraction index attain the form

$$n^2(z, \mathbf{x}) - 1 = \sigma_c \, \mu \left(\frac{z}{l}, \frac{\mathbf{x}}{l} \right) . \tag{2.8}$$

We assume that the fluctuations are modeled by an isotropic and smooth in $\mathbf{x}$, zero mean, stationary rapidly decorrelating random field $\mu(\cdot, \cdot)$, which moreover is Markovian in z. The normalized and dimensionless covariance is given by

$$R(\Delta z, \Delta \mathbf{x}) = \mathbb{E}[\mu(z', \mathbf{x}')\mu(\Delta z + z', \Delta \mathbf{x} + \mathbf{x}')] , \tag{2.9}$$

with $R(0, 0) = 1$. We, again, assume rapidly decaying correlations. Note that the dimensionless function R is supported on the $\mathcal{O}(1)$ scale. Thus, the correlation radius of the medium fluctuations is l. We shall consider a specific scaling regime characterized by a certain relation between the parameters that we have introduced in the problem; this scaling regime essentially corresponds to the one introduced in [33]. The regime will follow from the next step where we introduce dimensionless coordinates. To this end, we introduce the characteristic length scales as

- l_x, the characteristic length scale in the lateral direction,
- l_z, the characteristic length scale in the principal (depth) direction.

The following dimensionless parameters will be important in the further analysis:

$$\varepsilon = \frac{l}{l_z} , \quad \delta = \frac{l}{l_x} , \quad \theta = \frac{k_0 l_x^2}{l_z} ; \tag{2.10}$$

θ is commonly referred to as the *Rayleigh* number, while ε and δ are the medium correlation length relative to, respectively, characteristic principal and lateral scales. We remark that we specify in (3.5) below how the correlation scale of the ambient noise field relates to these small parameters. The important scaling regime considered here is the *high-frequency paraxial scaling* as introduced in [33], with

$$\frac{1}{\theta} \ll \varepsilon \ll \delta \ll 1 . \tag{2.11}$$

We hasten to add that there are other relevant scaling regimes [2, 5, 18, 34]; we discuss some scaling alternatives in section 5. However, the regime set forth above is characteristic for the estimation problem at hand and captures key aspects of the physical phenomenon under discussion. Note that it follows from our scaling assumptions that $l_x \ll l_z$, which corresponds to the characteristic propagation distance being much larger than the characteristic aperture size and the classical paraxial or beam scaling. This follows, since with $\vec{x}_j = (z_j, \mathbf{x}_j)$ denoting the recording points as before, we shall assume a regime where

$$z_j' = \frac{z_j}{l_z} , \quad \mathbf{x}_j' = \frac{\mathbf{x}_j}{l_x} ,$$

as well as

$$A' = \frac{A}{l_x} , \quad z_{\max}' = \frac{z_{\max}}{l_z} ,$$

are fixed, while considering the limits $1/\theta, \varepsilon, \delta \to 0$. The dimensionless coordinates are

$$z' = \frac{z}{l_z} , \quad \mathbf{x}' = \frac{\mathbf{x}}{l_x} , \quad k' = \frac{k}{k_0} , \quad \omega' = \frac{\omega}{\omega_0} , \tag{2.12}$$

and we let

$$t' = \frac{2\pi t}{T_0} = \omega_0 t\,, \quad \bar{c} = \frac{c_0}{\omega_0 l_z} = \frac{1}{k_0 l_z}\,, \tag{2.13}$$

so that

$$\omega(z/c_0 - t) = \omega'(z'/\bar{c} - t')\,.$$

Note that we have $k' = \omega'$.

In the further analysis, we shall drop the primes in the nondimensionalized coordinates; in dimensionless variables $(z, \mathbf{x}, \omega)$, the paraxial wave equation then becomes

$$2\mathrm{i}\,(\theta k)\,\frac{\partial \psi}{\partial z} + \triangle_{\mathbf{x}}\psi + \frac{\delta}{\sqrt{\varepsilon}}\,(\theta^2 k^2)\,\mu\Big(\frac{z}{\varepsilon}, \frac{\mathbf{x}}{\delta}\Big)\,\psi = 0\,, \tag{2.14}$$

with $\psi = \psi(z, \mathbf{x}, k)$, upon the identification (cf. (2.7))

$$\sigma_c = \sqrt{\varepsilon}\left(\frac{\varepsilon}{\delta}\right) = \sqrt{l/l_z}\left(\frac{l_x}{l_z}\right)\,.$$

This is the product of the white noise normalization factor $\sqrt{\varepsilon}$ and the paraxial scaling parameter. These factors are small so that the modeling corresponds to relatively small medium fluctuations. This is exactly the scaling of the fluctuations that gives partly coherent propagation of the wavefield. The fluctuations are sufficiently strong so that the wavefield is affected beyond the homogenization situation, but not so strong that the field completely loses its coherence.

In the scaling regime considered, $\frac{\partial^2 \psi}{\partial z^2}$ on the right-hand side of (2.7) can be neglected. Equation (2.14) is an evolution equation, which is supplemented with the initial conditions (cf. (2.2))

$$\frac{1}{2\pi}\int e^{\mathrm{i}\omega(-t)}\psi(z = 0, \mathbf{x}, \omega)\,d\omega = Y\,(T_0 t, l_x \mathbf{x})\,. \tag{2.15}$$

Indeed, the paraxial field ψ satisfies an initial value problem rather than a boundary value problem as in the case of the Helmholtz equation. This reflects the fact that we consider a regime where lateral scattering is dominant over scattering along the principal direction.

2.3. White noise model. We shall consider functionals of the field, ψ, in the scaling regime in (2.11). Following [33], we introduce the Wigner distribution:

$$W_\theta(z, \mathbf{x}, \mathbf{p}; \omega) = \iint \frac{1}{(2\pi)^d} e^{\mathrm{i}\mathbf{p}\cdot\mathbf{y}}\psi\left(z, \mathbf{x} - \frac{\mathbf{y}}{2\theta}, \omega\right)\psi\left(z, \mathbf{x} + \frac{\mathbf{y}}{2\theta}, \omega\right)^* d\mathbf{y}\,, \tag{2.16}$$

where we have used "*" to represent complex conjugation. It can then be shown [17, 33] that in the *high-frequency* ($\theta \to \infty$) and *white noise* ($\varepsilon \to 0$) limit, the limiting Wigner transform that we denote as W_δ is characterized weakly (in law) by the Itô–Liouville stochastic partial differential equation, as follows.

PROPOSITION 2.1. *The Wigner distribution W_θ converges in the limit $1/\theta \to 0$ followed by $\varepsilon \to 0$ weakly in law to the process W_δ solving*

$$dW_\delta = \left[-\frac{\mathbf{p}}{k}\cdot\nabla_{\mathbf{x}}W_\delta + \frac{k^2 D}{2}\triangle_{\mathbf{p}}W_\delta\right]dz - \frac{k}{2}\nabla_{\mathbf{p}}W_\delta \cdot d\mathbf{B}\left(z, \frac{\mathbf{x}}{\delta}\right)\,. \tag{2.17}$$

Here, $\mathbf{B}(\mathbf{x}, z)$ is a vector-valued Brownian field with covariance

$$\mathbb{E}[B_i(z_1, \mathbf{x}_1)B_j(z_2, \mathbf{x}_2)] = -\left(\frac{\partial^2 R_0(\mathbf{x}_1 - \mathbf{x}_2)}{\partial x_i \partial x_j}\right) z_1 \wedge z_2\,, \tag{2.18}$$

where $z_1 \wedge z_2 = \min\{z_1, z_2\}$; in the assumed isotropic case, we have (cf. (2.9))

$$D = -\frac{R_0''(0)}{4}, \quad \text{with } R_0(\mathbf{x}) = \int_{-\infty}^{\infty} R(\zeta, \mathbf{x})\, d\zeta\,. \tag{2.19}$$

We remark that the second derivative of the (isotropic) medium correlation function is negative: $R_0''(0) < 0$ so that indeed (2.17) is well posed. It now follows directly from (2.17) that the mean, $\overline{W} = \mathbb{E}[W_\delta]$, is independent of δ and solves the advection-diffusion equation

$$\frac{\partial \overline{W}}{\partial z} + \frac{\mathbf{p}}{k} \cdot \nabla_{\mathbf{x}} \overline{W} = \frac{k^2 D}{2} \triangle_{\mathbf{p}} \overline{W}\,. \tag{2.20}$$

The explicit expression for the Green's function of (2.20) is

$$\begin{aligned} U(z, \mathbf{x}, \mathbf{p}; \mathbf{x}_0, \mathbf{p}_0) = \iint \frac{1}{\omega^d (2\pi)^{2d}} \exp\left(\mathrm{i}\left[\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}_0) + \mathbf{r} \cdot \frac{(\mathbf{p} - \mathbf{p}_0)}{\omega} - z\mathbf{w} \cdot \frac{\mathbf{p}_0}{\omega}\right]\right) \\ \times \exp\left(-\frac{Dz}{2}\left[r^2 + z\mathbf{r} \cdot \mathbf{w} + \frac{w^2 z^2}{3}\right]\right) d\mathbf{w}\, d\mathbf{r}\,, \end{aligned} \tag{2.21}$$

with $r = \|\mathbf{r}\|_2$ and $w = \|\mathbf{w}\|_2$. This expression will be useful in order to characterize how the computed cross-correlations relate to the propagation Green's function of interest. That the computed cross-correlations give a stable and "low noise" estimate of the Green's function shall emerge as a consequence of assuming a stabilization regime. Indeed, here we shall assume the *stabilization* regime corresponding to the limit $\delta \to 0$, as discussed in [33]. The robust estimation of the empirical Green's function will be a consequence of the following stabilization result.

PROPOSITION 2.2. *Assume that the initial Wigner distribution, $W_I(\mathbf{x}, \mathbf{p})$, is uniformly bounded and Lipschitz continuous. Define*

$$I_\delta(z, \mathbf{x}, \mathbf{y}) = \iint W_\delta(z, \mathbf{x}, \mathbf{p}) e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}\, d\mathbf{p}. \tag{2.22}$$

Then

$$\lim_{\delta \to 0} \mathbb{E}\left\{I_\delta^2(z, \mathbf{x}, \mathbf{y})\right\} = \mathbb{E}^2\left\{I_\delta(z, \mathbf{x}, \mathbf{y})\right\}, \tag{2.23}$$

where $\mathbb{E}\{I_\delta(z, \mathbf{x}, \mathbf{y})\}$ is independent of δ.

This result is a slight generalization of the stability result derived in [33]; see the appendix.

3. Analysis of cross-correlations.

3.1. Time averaging.

The quantity of interest is the cross-correlation function,

$$\mathcal{C}_{\mathcal{H}}(\tau, \vec{x}_1, \vec{x}_2) = \int \mathcal{H}(t)\, u(t, \vec{x}_1) u(t + \tau, \vec{x}_2)\, dt\,, \tag{3.1}$$

in which $\mathcal{H}$ is a time-window function; cf. (1.1). Here, u is modeled from the solution of the paraxial wave equation (cf. (2.14)) subject to initial conditions (2.15). We introduce the notation $\check{v}$ for the partial Fourier transform of v with respect to time, and $\widehat{v}$ for the complete transform:

$$v(t,\mathbf{x}) = \frac{1}{2\pi}\int e^{-\mathrm{i}\omega t}\check{v}(\omega,\mathbf{x})\,d\omega = \frac{1}{(2\pi)^{d+1}}\iint e^{-\mathrm{i}(\omega t-\mathbf{p}\cdot\mathbf{x})}\widehat{v}(\omega,\mathbf{p})\,d\mathbf{p}\,d\omega\,.$$

Let G_θ be the Green's function associated with the paraxial wave equation (2.14); then we have in the standardized variables

$$u(t,\vec{x}) = \iint \widetilde{G}_\theta(t-s,\vec{x};\vec{x}_n)\,Y(T_0 s, l_x\mathbf{x}_n)\,d\mathbf{x}_n ds, \tag{3.2}$$

where $\vec{x}_n = (0,\mathbf{x}_n)$ and

$$\widetilde{G}_\theta(t,z,\mathbf{x};0,\mathbf{x}_n) = \frac{1}{2\pi}\int e^{\mathrm{i}\omega(z/\overline{c}-t)}G_\theta(\omega,z,\mathbf{x};0,\mathbf{x}_n)\,d\omega \tag{3.3}$$

(cf. (2.6)). Substituting (3.2) into (3.1) yields

$$\begin{aligned}\mathcal{C}_{\mathcal{H}}(\tau,\vec{x}_1,\vec{x}_2) = \iint &\mathcal{H}(t)\widetilde{G}_\theta(t-s_1,\vec{x}_1;\vec{x}_{n_1})Y(T_0 s_1, l_x\mathbf{x}_{n_1})\\ &\times \widetilde{G}_\theta(t+\tau-s_2,\vec{x}_2;\vec{x}_{n_2})Y(T_0 s_2, l_x\mathbf{x}_{n_2})\,dt\,ds_1\,ds_2\,d\mathbf{x}_{n_1}\,d\mathbf{x}_{n_2}\,.\end{aligned} \tag{3.4}$$

Note that by the result (2.20) the paraxial field decorrelates laterally on the scale $1/\theta$. We now assume that the ambient noise field decorrelates on this scale by choosing

$$X_0 = \frac{l_x}{\theta}\,. \tag{3.5}$$

When we substitute (2.2) into (3.4), we then obtain

$$\begin{aligned}\mathcal{C}_{\mathcal{H}}(\tau,\vec{x}_1,\vec{x}_2) = \iint \sigma_y^2 \left\{\int \mathcal{H}(t)\nu(t-v_1,\theta\mathbf{x}_{n_1})\,\nu(t-v_2,\theta\mathbf{x}_{n_2})\;dt\right\}\\ \times\,\widetilde{G}_\theta(v_1,\vec{x}_1;\vec{x}_{n_1})\widetilde{G}_\theta(v_2+\tau,\vec{x}_2;\vec{x}_{n_2})\chi\left(\frac{\mathbf{x}_{n_1}-\mathbf{x}_c}{A}\right)\chi\left(\frac{\mathbf{x}_{n_2}-\mathbf{x}_c}{A}\right)\,dv_1\,dv_2\,d\mathbf{x}_{n_1}\,d\mathbf{x}_{n_2}\,.\end{aligned} \tag{3.6}$$

We may choose for $\mathcal{H}$ the indicator function

$$\mathcal{H}_T(\cdot) = \frac{\mathcal{I}_{\left(-\frac{T}{2},\frac{T}{2}\right)}(\cdot)}{T}\,;$$

then

$$\lim_{T\to\infty}\int_{-\infty}^{\infty}\mathcal{H}_T(t)\nu(t-v_1,\theta\mathbf{x}_{n_1})\,\nu(t-v_2,\theta\mathbf{x}_{n_2})\,dt = C_0(v_2-v_1,\theta(\mathbf{x}_{n_2}-\mathbf{x}_{n_1}))\,,$$

the mean square with respect to the distribution of the impinging noise sources. Our interest is in such a regime where the time average effectively removes the fluctuations in the quantity of interest, *exploiting the randomness of the sources.* Substituting the above average into (3.6) leads to the introduction of

$$\begin{aligned}\langle\mathcal{C}\rangle_\theta(\tau,\vec{x}_1,\vec{x}_2) = \iint \sigma_y^2 C_0(v_2-v_1,\theta(\mathbf{x}_{n_2}-\mathbf{x}_{n_1}))\\ \times\,\widetilde{G}_\theta(v_1,\vec{x}_1;\vec{x}_{n_1})\widetilde{G}_\theta(v_2+\tau,\vec{x}_2;\vec{x}_{n_2})\chi\left(\frac{\mathbf{x}_{n_1}-\mathbf{x}_c}{A}\right)\chi\left(\frac{\mathbf{x}_{n_2}-\mathbf{x}_c}{A}\right)\,dv_1\,dv_2\,d\mathbf{x}_{n_1}\,d\mathbf{x}_{n_2}\,.\end{aligned} \tag{3.7}$$

3.2. Green's function filter and limits. We obtain the following representation for the quantity of interest.

PROPOSITION 3.1. *Let $\langle \mathcal{C} \rangle_\theta$ be as in* (3.7), *and assume the relative ordering*

$$z_1 < z_2 \,.$$

Then

$$\langle \mathcal{C} \rangle_\theta (\tau, \vec{x}_1, \vec{x}_2) \;=\; \iint \widetilde{G}_\theta\Big(\tau - s, z_2, \mathbf{x}_2; z_1, \mathbf{x}_1 - \frac{\mathbf{y}}{\theta}\Big)\, \Lambda_\theta\Big(z_1, s, \mathbf{x}_1 - \frac{\mathbf{y}}{2\theta}, \mathbf{y}\Big)\; ds\, d\mathbf{y}, \tag{3.8}$$

in which

$$\begin{aligned}\Lambda_\theta(z, \tau, \mathbf{x}, \mathbf{y}) = \int \frac{e^{-\mathrm{i}\omega\tau}}{2\pi} \Big\{ \iint & \sigma_y^2 \check{C}_0(\omega, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})) \\ & \times G_\theta\Big(\omega, z, \mathbf{x} - \frac{\mathbf{y}}{2\theta}; 0, \mathbf{x}_{n_2}\Big)\, G_\theta\Big(\omega, z, \mathbf{x} + \frac{\mathbf{y}}{2\theta}; 0, \mathbf{x}_{n_1}\Big)^* \\ & \times \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right)\, d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2} \Big\}\, \theta^{-d}\, d\omega \,.\end{aligned} \tag{3.9}$$

Here, Λ_θ is referred to as the Green's function filter.

Proof. We begin with substituting (3.3) into (3.7), and we obtain

$$\begin{aligned}\langle \mathcal{C} \rangle_\theta (\tau, \vec{x}_1, \vec{x}_2) = (2\pi)^{-2} & \iint \sigma_y^2 C_0(v_2 - v_1, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})) \\ \times\; G_\theta(\omega_1, z_1, \mathbf{x}_1; 0, \mathbf{x}_{n_1})^* e^{\mathrm{i}\omega_1(v_1 - z_1/\overline{c})} & \,G_\theta(\omega_2, z_2, \mathbf{x}_2; 0, \mathbf{x}_{n_2}) e^{-\mathrm{i}\omega_2((v_2+\tau)-z_2/\overline{c})} \\ \times\; \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right) & dv_1\, dv_2\, d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2}\, d\omega_1\, d\omega_2 \,,\end{aligned} \tag{3.10}$$

where we made use of the fact that $\widetilde{G}_\theta$ is real-valued. We carry out the integration over v_1 yielding a Fourier transform of C_0 with respect to its time argument; the integration over v_2 then gives a factor $\delta(\omega_2 - \omega_1)$:

$$\begin{aligned}\langle \mathcal{C} \rangle_\theta (\tau, \vec{x}_1, \vec{x}_2) &= (2\pi)^{-2} \iint \sigma_y^2 \check{C}_0(\omega_1, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})) \\ &\quad\times G_\theta(\omega_1, z_1, \mathbf{x}_1; 0, \mathbf{x}_{n_1})^* G_\theta(\omega_2, z_2, \mathbf{x}_2; 0, \mathbf{x}_{n_2})\, e^{-\mathrm{i}((\omega_2-\omega_1)v_2+\omega_2\tau)} e^{\mathrm{i}(\omega_2 z_2 - \omega_1 z_1)/\overline{c}} \\ &\quad\times \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right)\, dv_2\, d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2}\, d\omega_1\, d\omega_2 \\ &= (2\pi)^{-1} \iint \sigma_y^2 \check{C}_0\left(\omega, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})\right)\, e^{-\mathrm{i}\omega(\tau - (z_2 - z_1)/\overline{c})} \\ &\quad\times G_\theta(\omega, z_1, \mathbf{x}_1; 0, \mathbf{x}_{n_1})^* G_\theta(\omega, z_2, \mathbf{x}_2; 0, \mathbf{x}_{n_2}) \\ &\quad\times \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right)\, d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2}\, d\omega \,.\end{aligned} \tag{3.11}$$

We invoke the semigroup property of the solution operator to the paraxial wave equation, and, using that $z_2 > z_1$, we get

$$\begin{aligned}(3.12)\quad \left\langle \mathcal{C} \right\rangle_\theta (\tau, \vec{x}_1, \vec{x}_2) = \iint \frac{e^{-\mathrm{i}\omega(\tau-(z_2-z_1)/\bar{c})}}{2\pi} \Bigg\{ \iint \sigma_y^2 \check{C}_0(\omega, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})) \\ \times\, G_\theta(\omega, z_1, \mathbf{x}_1; 0, \mathbf{x}_{n_1})^* G_\theta(\omega, z_1, \mathbf{y}; 0, \mathbf{x}_{n_2})\, \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right) d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2} \Bigg\} \\ \times\, G_\theta(\omega, z_2, \mathbf{x}_2; z_1, \mathbf{y})\, d\omega\, d\mathbf{y}\,.\end{aligned}$$

We then change variables of integration,

$$\begin{aligned}(3.13)\quad \left\langle \mathcal{C} \right\rangle_\theta (\tau, \vec{x}_1, \vec{x}_2) = \iint \frac{e^{-\mathrm{i}\omega(\tau-(z_2-z_1)/\bar{c})}}{2\pi} \Bigg\{ \iint \sigma_y^2 \check{C}_0(\omega, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1})) \\ \times\, G_\theta(\omega, z_1, \mathbf{x}_1; 0, \mathbf{x}_{n_1})^* G_\theta\Big(\omega, z_1, \mathbf{x}_1 - \frac{\mathbf{y}}{\theta}; 0, \mathbf{x}_{n_2}\Big) \\ \times\, \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right) d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2} \Bigg\} G_\theta\Big(\omega, z_2, \mathbf{x}_2; z_1, \mathbf{x}_1 - \frac{\mathbf{y}}{\theta}\Big)\, \theta^{-d}\, d\omega\, d\mathbf{y} \\ = \iint \frac{e^{-\mathrm{i}\omega(\tau-(z_2-z_1)/\bar{c})}}{2\pi} G_\theta\Big(\omega, z_2, \mathbf{x}_2; z_1, \mathbf{x}_1 - \frac{\mathbf{y}}{\theta}\Big)\, \check{\Lambda}_\theta\Big(z_1, \omega, \mathbf{x}_1 - \frac{\mathbf{y}}{2\theta}, \mathbf{y}\Big)\, d\omega\, d\mathbf{y}\,,\end{aligned}$$

which gives the result (3.8). □

We will now make the following assumption.

Assumption 1.

$$(3.14)\qquad \sigma_y^2 = \theta^d\,.$$

The expression (3.9) then simplifies, and this assumption corresponds to letting the correlation radius of the impinging noise field be θ independent.

We remark that the paraxial Green's function will decorrelate in the lateral dimensions on the $1/\theta$ scale, and that this is the motivation for the choice of parameterization of the filter Λ_θ.

It will prove natural to introduce the Wigner distribution

$$\begin{aligned}(3.15)\quad W_\theta(z, \mathbf{x}, \mathbf{p}; \omega) = \iint \frac{e^{\mathrm{i}\mathbf{p}\cdot\mathbf{y}}}{(2\pi)^d}\, G_\theta\Big(\omega, z, \mathbf{x} - \frac{\mathbf{y}}{2\theta}; 0, \mathbf{x}_{n_2}\Big)\, G_\theta\Big(\omega, z, \mathbf{x} + \frac{\mathbf{y}}{2\theta}; 0, \mathbf{x}_{n_1}\Big)^* \\ \times\, \check{C}_0(\omega, \theta(\mathbf{x}_{n_2} - \mathbf{x}_{n_1}))\, \chi\left(\frac{\mathbf{x}_{n_1} - \mathbf{x}_c}{A}\right) \chi\left(\frac{\mathbf{x}_{n_2} - \mathbf{x}_c}{A}\right) d\mathbf{x}_{n_1}\, d\mathbf{x}_{n_2}\, d\mathbf{y}\,,\end{aligned}$$

so that

$$(3.16)\qquad \check{\Lambda}_\theta(z, \omega, \mathbf{x}, \mathbf{y}) = \iint W_\theta(z, \mathbf{x}, \mathbf{p}; \omega) e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}} d\mathbf{p}\,.$$

This Wigner distribution coincides with the one given in (2.16) subject to initial conditions that derive from (2.15), which follows from (2.6), (3.2), and (3.3). In the high-frequency ($\theta \to \infty$) and white noise ($\varepsilon \to 0$) limits, the Wigner distribution in (3.15) is characterized weakly (in law) by (2.17), while its mean satisfies (2.20).

The initial condition, at $z = 0$, for the Wigner distribution follows directly from the corresponding initial condition for the Green's function G_θ:

$$
\begin{aligned}
(3.17)\quad W_\theta(0,\mathbf{x},\mathbf{p};\omega) = \iint & \frac{e^{\mathrm{i}\theta\mathbf{p}\cdot 2(\mathbf{x}-\mathbf{x}_{n_2})}}{(2\pi)^d}\check{C}_0(\omega,\theta(2(\mathbf{x}_{n_2}-\mathbf{x})))\\
& \times \chi\left(\frac{2\mathbf{x}-\mathbf{x}_{n_2}-\mathbf{x}_c}{A}\right)\chi\left(\frac{\mathbf{x}_{n_2}-\mathbf{x}_c}{A}\right)\theta^d\, d\mathbf{x}_{n_2}\\
& \sim \widehat{C}_0\left(\omega,\mathbf{p}\right)\frac{\chi^2\left(\frac{\mathbf{x}-\mathbf{x}_c}{A}\right)}{(4\pi)^d} \equiv W_I(\mathbf{x},\mathbf{p};\omega) \quad \text{as } \theta\to\infty.
\end{aligned}
$$

We will apply this approximation below.

We now use the Green's function (2.21), and initial condition (3.17), in (3.16) and define the following Green's function filter,

$$
\begin{aligned}
\check{\Lambda}(z,\omega,\mathbf{x},\mathbf{y}) = & \iint \frac{e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}}{(2\pi)^{2d}}\exp\left(\mathrm{i}\left[\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_0)+\mathbf{r}\cdot\frac{(\mathbf{p}-\mathbf{p}_0)}{\omega}-z\mathbf{w}\cdot\frac{\mathbf{p}_0}{\omega}\right]\right)\\
& \times \exp\left(-\frac{Dz}{2}\left[r^2+z\mathbf{r}\cdot\mathbf{w}+\frac{w^2z^2}{3}\right]\right)\widehat{C}_0\left(\omega,\mathbf{p}_0\right)\\
& \times \frac{\chi^2\left(\frac{\mathbf{x}_0-\mathbf{x}_c}{A}\right)}{(4\pi)^d}\omega^{-d}\, d\mathbf{w}\, d\mathbf{r}\, d\mathbf{p}\, d\mathbf{x}_0\, d\mathbf{p}_0\\
= & \frac{1}{(4\pi)^d}\iint e^{\mathrm{i}\omega\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_c)}\exp\left(-\frac{\omega^2Dz}{2}\left[y^2+z\mathbf{y}\cdot\mathbf{w}+\frac{w^2z^2}{3}\right]\right)\\
(3.18)\qquad & \times\ \check{C}_0(\omega,\mathbf{y}+z\mathbf{w})\widehat{\chi}_A\left(\omega\mathbf{w}\right)\omega^d\, d\mathbf{w}\,,
\end{aligned}
$$

in which

$$
\widehat{\chi}_A\left(\mathbf{w}\right) = \iint e^{-\mathrm{i}\mathbf{w}\cdot\mathbf{x}_0}\,\chi^2\left(\frac{\mathbf{x}_0}{A}\right)d\mathbf{x}_0\,.
$$

The characterization of a statistically stable filter now follows from the representation (3.16) and Propositions 2.1 and 2.2, as follows.

Proposition 3.2. *The deterministic Green's function filter converges to a deterministic filter*

$$
(3.19)\qquad \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{\theta\to\infty}\check{\Lambda}_\theta\left(z,\omega,\mathbf{x},\mathbf{y}\right) = \check{\Lambda}(z,\omega,\mathbf{x},\mathbf{y})
$$

in $L^2(\mathbb{P})$.

3.3. Effective filter. In order to obtain qualitative and quantitative insight on the Green's function filtering behavior, we shall assume that the spectrum of the envelope function and the spectrum of the noise covariance have Gaussian shapes. That is, we shall assume

$$
(3.20)\qquad \chi(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}}e^{-\frac{|\mathbf{x}|^2}{2}}\,, \quad \check{C}_0\left(\omega,\mathbf{y}\right) = \widehat{f}_0(\omega)e^{\frac{-|\mathbf{y}|^2}{2\sigma_x^2}}\,.
$$

We recall that the filter will be evaluated at $z = z_1$, corresponding to the longitudinal distance from the source plane to the first recording point. We find that then

$$
\begin{aligned}
\check{\Lambda}(z,\omega,\mathbf{x},\mathbf{y}) = & \left(\frac{A}{4\pi\sqrt{2}}\right)^d\omega^d\widehat{f}_0(\omega)\iint e^{\mathrm{i}\omega\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_c)}\exp\left(-\frac{\omega^2Dz}{2}\left[y^2+z\mathbf{y}\cdot\mathbf{w}+\frac{w^2z^2}{3}\right]\right)\\
& \times e^{\frac{-|\mathbf{y}+z\mathbf{w}|^2}{2\sigma_x^2}}e^{\frac{-\omega^2|\mathbf{w}|^2A^2}{4}}\, d\mathbf{w}\,.
\end{aligned}
$$

We rewrite this expression as

(3.21)

$$\begin{aligned}\check{\Lambda}(z,\omega,\mathbf{x},\mathbf{y}) &= \left(\frac{A}{4\pi\sqrt{2}}\right)^{d} \omega^{d}\widehat{f}_0(\omega)\iint e^{\mathrm{i}\omega\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_c)}e^{-\omega^2\left(a_1 w^2+2a_2\mathbf{y}\cdot\mathbf{w}+a_3 y^2\right)/2}\,d\mathbf{w}\\ &= \left(\frac{A^2}{16\pi a_1}\right)^{d/2}\widehat{f}_0(\omega)e^{-\frac{|\mathbf{x}-\mathbf{x}_c|^2}{2a_1}}e^{-\omega^2(a_3-a_2^2/a_1)y^2/2}e^{-\mathrm{i}\omega(a_2/a_1)\mathbf{y}\cdot(\mathbf{x}-\mathbf{x}_c)}\end{aligned}$$

with

$$a_1(\omega)=\frac{A^2}{2}+\frac{z^2}{\omega^2\sigma_x^2}+\frac{Dz^3}{3},\quad a_2(\omega)=\frac{z}{\omega^2\sigma_x^2}+\frac{Dz^2}{2},\quad a_3(\omega)=\frac{1}{\omega^2\sigma_x^2}+Dz\,.$$

Furthermore, we shall assume that the noise source field has a temporal frequency spectrum of the form

$$\widehat{f}_0(\omega)=\frac{1}{2}\left(\widehat{f}\,(T_s(\omega-\omega_c))+\widehat{f}\,(T_s(\omega+\omega_c))\right)\,. \tag{3.22}$$

We can now associate two characteristic length scales and one characteristic temporal scale with the Green's function filter in the narrow band situation so that T_s is large:

- The *refocusing* length scale

$$\begin{aligned}l_f(\omega_c) &= \frac{\lambda_c}{2\pi\sqrt{a_3(\omega_c)-\frac{a_2^2(\omega_c)}{a_1(\omega_c)}}}\\ &= \frac{\lambda_c}{2\pi}\left(\frac{\frac{A^2}{2}+\frac{\lambda_c^2 z^2}{4\pi^2\sigma_x^2}+\frac{Dz^3}{3}}{\left(A^2+\frac{Dz^3}{6}\right)\frac{Dz}{2}+\left(\frac{A^2}{2}+\frac{Dz^3}{3}\right)\frac{\lambda_c^2}{4\pi^2\sigma_x^2}}\right)^{1/2},\end{aligned}$$

  for

$$\lambda_c = 2\pi/\omega_c$$

  with ω_c a characteristic frequency of the noise source spectrum. This length scale determines the smoothing scale in the spatial source coordinates of the Green's function estimate, since the Green's function is blurred on the scale l_f/θ; see (3.19), (3.21). Note that in the low-frequency limit with $\lambda_c \to \infty$ the estimate of the Green's function degrades since the refocusing length scale becomes large in this limit.
- The *effective aperture* length scale

$$l_a(\omega_c)=\sqrt{a_1(\omega_c)}=\sqrt{\frac{A^2}{2}+\frac{\lambda_c^2 z^2}{4\pi^2\sigma_x^2}+\frac{Dz^3}{3}}\,.$$

  This corresponds to the lateral range, relative to the center $\mathbf{x}_c$, of impinging noise sources contributing to the Green's function estimate through $\langle\mathcal{C}\rangle_\theta$. If the lateral separation between $\mathbf{x}_j$ and $\mathbf{x}_c$ is large relative to this length, then the filter will weaken the Green's function significantly. In the limit $\lambda_c \to 0$, l_a corresponds to the effective aperture in [33].

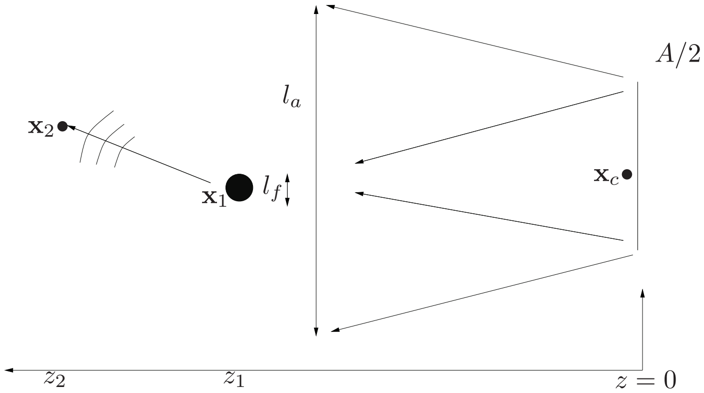


FIG. 3.1. *The effective aperture l_a and the refocusing length scale l_f.*

- The Green's function will be blurred in time, on the scale T_s, corresponding to the support of the ambient noise correlations in time.

The characteristic length scales are illustrated in Figure 3.1.

We comment on how the characteristic length scales depend on some of the parameters. In the case of a homogeneous medium, with $D = 0$ and low frequencies, we have

$$l_f(\omega_c)|_{D=0} \overset{\lambda_c \to \infty}{\sim} \left(\frac{\lambda_c z}{A}\right) \frac{1}{\sqrt{2}\pi} \,,$$

which corresponds to the classical *Rayleigh* resolution. While in the high-frequency limit and homogeneous medium case, with $D = 0$, we find

$$l_f(\omega_c)|_{D=0} \overset{\lambda_c \to 0}{\sim} \sigma_x \,,$$

that is, a length scale corresponding to that of the support of the ambient noise field in the lateral spatial dimensions. Next, we consider the limit of (relatively) strong medium fluctuations:

$$l_f(\omega_c) \overset{D \to \infty}{\sim} \frac{\lambda_c}{\pi\sqrt{zD}} \,,$$

which leads to a small refocusing scale.

4. Source location estimation. Here, we consider the case of narrow noise aperture, A, and discuss the problem of estimating the "source location" $\mathbf{x}_c$. We assume that the points of observation lie in the plane $z_2 = z_1 = z$. From (3.8) we then find that

$$\langle \mathcal{C} \rangle_\theta (\tau, z, \mathbf{x}_1, z, \mathbf{x}_2) = \theta^d \int \frac{e^{-\mathrm{i}\omega\tau}}{2\pi} \check{\Lambda}_\theta(z, \omega, \tfrac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2), \theta(\mathbf{x}_1 - \mathbf{x}_2)) \, d\omega \,.$$

For two points separated on the $1/\theta$ scale, we write

$$\mathbf{x}_1 = \mathbf{x} + \frac{\mathbf{y}}{2\theta}, \quad \mathbf{x}_2 = \mathbf{x} - \frac{\mathbf{y}}{2\theta} \,. \tag{4.1}$$

In the case of source estimation, we replace assumption (3.14) by the following.

ASSUMPTION 2.

$$\sigma_y^2 \equiv 1 \,. \tag{4.2}$$

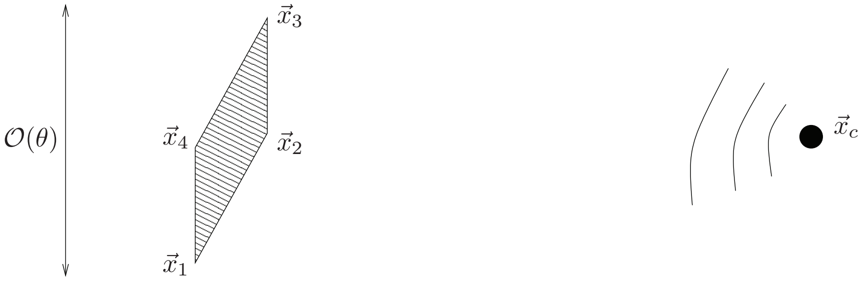


FIG. 4.1. *Estimation of source location: By computing the cross-correlations between the points* $\vec{x}_1, \ldots, \vec{x}_4$ *on the mirror, one obtains differential traveltimes, corresponding to* $|\vec{x}_i - \vec{x}_c| - |\vec{x}_j - \vec{x}_c|$, *which can be used to estimate the source location* $\vec{x}_c$.

Thus, the magnitude of the localized noise source field is now θ independent. We then have

$$\langle \mathcal{C} \rangle_\theta \Big(\tau, z, \mathbf{x} + \frac{\mathbf{y}}{2\theta}, z, \mathbf{x} - \frac{\mathbf{y}}{2\theta} \Big) = \int \frac{e^{-i\omega\tau}}{2\pi} \iint W_\theta(z, \mathbf{x}, \mathbf{p}; \omega) e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}} d\mathbf{p}\,,$$

and it follows by Proposition 2.2 that $\langle \mathcal{C} \rangle_\theta$ is statistically stable.

We consider, moreover, a tight lateral support for the noise field correlations: $\sigma_x \ll 1$ in the model (3.20). Then, using Proposition 3.2 and (3.18), we obtain

$$\begin{aligned} &\sigma_x^{-d} \langle \mathcal{C} \rangle_\theta \Big(\tau, z, \mathbf{x} + \frac{\mathbf{y}}{\theta}, z, \mathbf{x} - \frac{\mathbf{y}}{\theta} \Big) \\ &\qquad \sim \left(\frac{A^2}{16\pi z^2} \right)^{d/2} \int |\omega|^d \widehat{f}_0(\omega) \frac{e^{-\mathrm{i}\omega\tau}}{2\pi} e^{-\frac{\omega^2 \sigma_a^2 y^2}{2}} e^{-\mathrm{i}\omega \mathbf{y}\cdot(\mathbf{x}-\mathbf{x}_c)/z}\, d\omega \\ &\qquad = \left\{ \left(\frac{A^2}{16\pi z^2} \right)^{d/2} \widetilde{f}\,(\cdot) * \mathcal{N}_{\sigma_a}(\cdot) \right\} \left(\tau + \mathbf{y} \cdot \frac{\mathbf{x} - \mathbf{x}_c}{z} \right) \,, \end{aligned}$$

where "$*$" denotes convolution, $\mathcal{N}_\sigma$ is the Gaussian distribution with standard deviation σ, and we define

$$\begin{aligned} \sigma_a &= \sqrt{\frac{A^2}{2z^2} + \frac{Dz}{3}}\,, \\ \widetilde{f}(\tau) &= \int \frac{e^{-\mathrm{i}\omega\tau}}{2\pi} |\omega|^d \widehat{f}_0(\omega)\, d\omega. \end{aligned}$$

We remark that in the paraxial regime we have

$$\tau + \mathbf{y}\cdot(\mathbf{x} - \mathbf{x}_c)/z = \frac{2\pi}{T_0} \left\{ \widetilde{t} + \left(\frac{|\widetilde{\mathbf{x}}_1 - \widetilde{\mathbf{x}}_c|^2}{2\widetilde{z}c_0} - \frac{|\widetilde{\mathbf{x}}_2 - \widetilde{\mathbf{x}}_c|^2}{2\widetilde{z}c_0} \right) \right\} \approx \frac{2\pi}{T_0} (\widetilde{t} + (\widetilde{\tau}_{s1} - \widetilde{\tau}_{s2})\,, \tag{4.3}$$

with $\widetilde{t}, \widetilde{z}, \widetilde{\mathbf{x}}_j$, and $\widetilde{\mathbf{x}}_j$ denoting the original scaled coordinates and $\widetilde{\tau}_{sj}$ being the traveltime from the noise source at $\vec{x}_c$ to observation point $\vec{x}_j$. Thus, given observations at array points $\vec{x}_1, \ldots, \vec{x}_4$, separated as in (4.1), one can estimate the location of the source via differential traveltime estimates in the $(d+1) = 3$-dimensional case; see Figure 4.1. The resolution of the estimate is limited by the support of the noise correlation function, the strength of the medium fluctuations, and the aperture A. We remark that in contrast to the situation where one aims at estimating the Green's function, here, a large value for σ_a leads to a poor resolution.

5. White noise paraxial regimes and applications. In this section, we discuss alternative scaling scenarios that have been introduced in [34] and [20]. We demonstrate the robustness of the results derived above by showing that their essential features prevail under a wide range of scaling scenarios. Indeed, the general analysis will be the same, but the Green's function filter changes through replacing the function U in (2.21) by the Green's function associated with the relevant modification of (2.20). In subsection 5.4 we conclude the discussion on scaling scenarios by considering a particular application, namely, body-wave scattering in Southeastern Tibet motivated in the introduction, and its characteristic scales.

We consider a scaling and a nondimensionalization as above. We shall here discuss the situation with $\varepsilon \to 0$ being the smallest parameter. This limit gives the Itô form of (2.14):

$$2\mathrm{i}\,(\theta k)\,d_z\psi + \triangle_{\mathbf{x}}\psi\,dz + \frac{\mathrm{i}(\theta^3 k^3)\delta^2}{4} R_0(0)\psi\,dz + (\theta^2 k^2)\delta\psi d_z B\Big(z, \frac{\mathbf{x}}{\delta}\Big) = 0\,, \tag{5.1}$$

with the law of the Brownian flow ∇B coinciding with the law of the flow $\mathbf{B}$ in (2.17). This Itô form of the Schrödinger equation is discussed in, for instance, [11]. Now different regimes lead to different Wigner distributions and equations that they satisfy, and, hence, different Green's functions, which shape the Green's function filter.

5.1. Subsequent high-frequency scaling. First, we consider the situation with a subsequent high-frequency or geometrical optics limit followed by the large diversity scaling. That is, we have

$$\varepsilon \ll \frac{1}{\theta} \ll \delta \ll 1\,. \tag{5.2}$$

In this case, the Wigner distribution in (2.16) again satisfies (2.20)! Therefore, the Green's function estimation corresponds to the estimation discussed above.

5.2. The joint limit. We discuss next the joint limit with θ and δ going to zero simultaneously, that is,

$$\varepsilon \ll \frac{\xi}{\theta} = \delta \ll 1 \tag{5.3}$$

for $\xi = \mathcal{O}(1)$. As shown in [34], the Wigner distribution in (2.16) converges in the limit $\delta = \xi\theta \to 0$ weakly (in $\mathcal{S}'(\mathbb{R}^{2d})$) and in probability to $\widetilde{W}$, solving

$$\begin{aligned}\frac{\partial \widetilde{W}}{\partial z}(z,\mathbf{x},\mathbf{p}) &+ \frac{\mathbf{p}}{\omega}\cdot\nabla_{\mathbf{x}}\widetilde{W}(z,\mathbf{x},\mathbf{p}) \\ &= \frac{\omega^2\xi^2}{4}\iint_{-\infty}^{\infty}\frac{d\mathbf{q}}{(2\pi)^d}\widehat{R}_0(\mathbf{q})\left(\widetilde{W}(z,\mathbf{x},\mathbf{p}+\mathbf{q}) - \widetilde{W}(z,\mathbf{x},\mathbf{p})\right)\end{aligned} \tag{5.4}$$

(cf. (2.19) for the definition of R_0). The Green's function of (5.4) is explicitly given by

$$\begin{aligned}\widetilde{U}(z,\mathbf{x},\mathbf{p};\mathbf{x}^0,\mathbf{p}^0) = &\iint_{-\infty}^{\infty}\frac{1}{\omega^d(2\pi)^{2d}}\exp\left(\mathrm{i}\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}^0) + \mathrm{i}\mathbf{r}\cdot\frac{(\mathbf{p}-\mathbf{p}^0)}{\omega} - \mathrm{i}z\mathbf{w}\cdot\frac{\mathbf{p}^0}{\omega}\right) \\ &\times\exp\left(-\frac{\omega^2\xi^2}{4}\int_0^z D_R\left(\frac{\mathbf{r}+\mathbf{w}s}{\omega}\right)ds\right)d\mathbf{w}\,d\mathbf{r}\end{aligned} \tag{5.5}$$

and replaces U in (2.21). Here, D_R is the medium structure function,

$$D_R(\mathbf{r}) = R_0(\mathbf{0}) - R_0(\mathbf{r}); \tag{5.6}$$

cf. (2.19). The expression (3.18) for the Green's function filter then becomes

$$\begin{aligned}
\check{\tilde{\Lambda}}&(z,\omega,\mathbf{x},\mathbf{y};\sigma_t,\sigma_x) \\
&\approx \iint_{-\infty}^{\infty} \frac{e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}}{(2\pi)^{2d}} \exp\left(\mathrm{i}\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}^0) + \mathrm{i}\mathbf{r}\cdot\frac{(\mathbf{p}-\mathbf{p}^0)}{\omega} - \mathrm{i}z\mathbf{w}\cdot\frac{\mathbf{p}^0}{\omega}\right) \\
&\qquad \times \exp\left(-\frac{\omega^2\xi^2}{4}\int_0^z D_R\left(\frac{\mathbf{r}+s\mathbf{w}}{\omega}\right) ds\right) \\
&\qquad \times \widehat{C}_0\left(\omega,\mathbf{p}_0\right) \frac{\chi^2\left(\frac{\mathbf{x}_0-\mathbf{x}_c}{A}\right)}{(2\pi)^d} \left(\frac{1}{\omega}\right)^d d\mathbf{w}\, d\mathbf{r}\, d\mathbf{p}\, d\mathbf{x}_0\, d\mathbf{p}_0 \\
&= \frac{1}{(4\pi)^d} \iint e^{\mathrm{i}\omega\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_c)}\, \check{C}_0(\omega,\mathbf{y}+z\mathbf{w})\, e^{-\frac{\omega^2\xi^2}{4}\int_0^z D_R(\mathbf{y}+s\mathbf{w})\, ds}\, \widehat{\chi}_A\left(\omega\mathbf{w}\right) \omega^d\, d\mathbf{w}\,.
\end{aligned} \tag{5.7}$$

In the case where the correlation length associated with the structure function D_R is large, we can expand this, assumed smooth, function D_R,

$$D_R(\mathbf{y}) \mapsto 2D|\mathbf{y}|^2\,,$$

and recover (3.18). In the joint scaling limit, however, the whole spectrum of the medium fluctuations is involved in the definition of the Green's function filter.

5.3. Long-range media. We finally comment on the situation with rough and long-range media as, for instance, in the turbulent atmosphere and heterogeneous regions of the earth's crust. Here we will consider a white noise limit, however, with the Fresnel number and lateral diversity scales fixed. In this scaling, which is analyzed in detail by Fannjiang and Solna [19, 20, 21], (2.14) becomes

$$2\mathrm{i}\,(\theta k)\,\frac{\partial\psi}{\partial z} + \triangle_{\mathbf{x}}\psi + \frac{1}{\sqrt{\varepsilon}}\left(\theta^2 k^2\right)\widetilde{\mu}\left(\frac{z}{\varepsilon},\mathbf{x}\right)\psi = 0, \tag{5.8}$$

where the power spectrum of the random field, $\widetilde{\mu}(\cdot,\cdot)$, is given by

$$\Phi(\vec{\mathbf{k}}) \approx \sigma_H |\vec{\mathbf{k}}|^{-1-2H} |\vec{\mathbf{k}}|^{-d}\,, \tag{5.9}$$

for $|\vec{\mathbf{k}}|$ in the inertial range, where $\vec{\mathbf{k}} \in \mathbb{R}^{d+1}$ is the spectral variable and H the Hurst exponent characterizing the roughness of the medium fluctuations. That is, we assume that the medium fluctuations follow a power law form over a set of scales called the intertial range, which corresponds to turbulent or long-range medium modeling. In this case, the Wigner distribution solves, in the white noise limit, in the sense of L^2-weak solutions, a Wigner–Itô equation driven by an operator-valued Brownian motion. In particular, the first moment in (2.20) now becomes

$$\frac{\partial\overline{W}}{\partial z} + \frac{1}{\omega}\mathbf{p}\cdot\nabla_{\mathbf{x}}\overline{W} = \mathcal{Q}_0\overline{W}\,, \tag{5.10}$$

with

$$\mathcal{Q}_0\overline{W} = \frac{\theta^2\omega^2}{4}\int \Phi(0,\mathbf{q})\left(-2\overline{W}(\mathbf{p}) + \overline{W}\left(\mathbf{p}+\frac{\mathbf{q}}{\theta}\right) + \overline{W}\left(\mathbf{p}-\frac{\mathbf{q}}{\theta}\right)\right) d\mathbf{q};$$

see [20] for details. The Green's function of (5.10) is

$$\widetilde{U}_\theta(z,\mathbf{x},\mathbf{p};\mathbf{x}^0,\mathbf{p}^0) = \iint_{-\infty}^{\infty} \frac{1}{\omega^d(2\pi)^{2d}} \exp\left(\mathrm{i}\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}^0) + \mathrm{i}\mathbf{r}\cdot\frac{(\mathbf{p}-\mathbf{p}^0)}{\omega} - \mathrm{i}z\mathbf{w}\cdot\frac{\mathbf{p}^0}{\omega}\right)$$

$$\times \exp\left(-\frac{\omega^2\theta^2}{2}\int_0^z D_*\left(\frac{\mathbf{r}+\mathbf{w}s}{\theta\omega}\right)\right) d\mathbf{w}\, d\mathbf{r}\,, \tag{5.11}$$

corresponding to the form (5.5). The structure function can now be expressed as

$$D_*(\mathbf{x}) = \int \Phi(0,\mathbf{q})\left(1-e^{i\mathbf{x}\cdot\mathbf{q}}\right) d\mathbf{q}\,. \tag{5.12}$$

The Green's function filter is therefore again characterized by (5.7). For the power law medium we have the short distance asymptotic

$$D_*(\mathbf{x}) \approx C_*^2|\mathbf{x}|^{2H_*}\,, \tag{5.13}$$

where the effective Hölder exponent H_* is given by

$$H_* = \begin{cases} H+1/2 & \text{for } H\in(0,1/2), \\ 1 & \text{for } H\in(1/2,1]\,, \end{cases} \tag{5.14}$$

in which H is the Hölder exponent of the original medium and C_*^2 a structure parameter.

Note that the effective Hölder exponent H_* is always bigger than 1/2, corresponding to a "persistent" or a long-range power law. Using this asymptotic and considering a regime with relatively narrow support for the impinging noise field by setting

$$\check{C}_0\,(\sigma_t\omega,\mathbf{y}) \mapsto \widehat{f}_0\,(\omega)\delta(\mathbf{y})\,,$$

we find

$$\check{\Lambda}_H(z,\omega,\mathbf{x},\mathbf{y}) \approx \frac{\omega^d}{(4\pi z)^d} e^{-i\omega\mathbf{y}\cdot\mathbf{x}/z}\widehat{\chi}_A\left(\frac{-\omega\mathbf{y}}{z}\right)\widehat{f}_0\,(\omega)^*\, e^{-\frac{(\omega^2\theta^{2-2H_*}z|\mathbf{y}|^{2H_*})C_*^2}{2(2H_*+1)}}\,.$$

For the normalized noise field supported at the carrier frequency ω_c as in (3.22) we thus find that for a fixed Fresnel number the spatial support of the Green's function filter scales as ω_c^{-1/H_*}. That is, the resolution depends *nonlinearly* on the wavelength associated with the characteristic temporal scale of the impinging noise field. In the limit of rough media with $H_* \to 0$ the lateral spatial support of the Green's function filter scales like λ_c^2.

5.4. Applications. Here, we address the application of our analysis to passive seismic tomography, making use of continuous recordings in an array of detectors or receivers. We discuss which scaling regime would apply to the regional study in Southeastern Tibet [43]. In the latter study, the focus was on the "generation" of surface waves. Here, we seek insight into the behavior of body-wave contributions for future applications in the same region. Our analysis incorporates what seismologists refer to as "ambient noise" (our random source distribution) and "coda waves" (through random medium fluctuations).

Concerning the Southeastern Tibet data set discussed in section 1, we obtain the following characterization [43]. The characteristic distance between individual stations (receivers) is approximately $\mathcal{O}(100\text{km})$, and, similarly, the characteristic transversal distance between passive sources (magnitude $Mw > 5$) is $\mathcal{O}(100\text{km})$. (For smaller earthquakes—but with magnitude $Mw > 4$—the characteristic transversal distance reduces to $\mathcal{O}(10\text{km})$.) The distance from the array to the sources is $\mathcal{O}(5000\text{km})$; most of the sources are likely to be located in the Western Pacific margins and Eastern Indian Ocean margins. The dominant wavelength for shear (S) waves is $\mathcal{O}(20\text{km})$, while the dominant wavelength for compressional (P) waves is $\mathcal{O}(5\text{km})$. (Under certain simplifying conditions, shear waves have been modeled by a scalar wave equation, whence the current analysis would still be applicable.) The correlation length of medium fluctuations is, with the present knowledge, hard to estimate, but a value of $\mathcal{O}(10\text{km})$ is plausible, also, given the complexity in tectonics of the region. The "asymmetry" observed in the cross-correlations in [43] is explained and inherent in our setup based on the paraxial wave equation.

In our modeling, the characteristic transversal distance between stations or sources roughly corresponds to l_x, the distance from the array to the sources roughly corresponds to l_z, and the correlation length of medium fluctuations to l; the dominant wavelength is λ_0. For compressional body waves this results in $\delta \approx 10^{-1}$, $\varepsilon \approx 2\times 10^{-3}$, and $\theta^{-1} \approx 4 \times 10^{-1}$, thus, corresponding most closely to the scaling discussed in section 5.2.

6. Numerical illustrations. In this section we present a numerical illustration where we show the effect of the Green's function filter. We shall use the filter corresponding to the scaling regime discussed in section 3.3.

We assume that the medium is homogeneous for $z > z_1$ and plot the quantity

$$\mathcal{I}(\tau, \mathbf{x}_1; A, A_D) = \iint G_\theta\Big(\tau - s, z_2, \mathbf{x}_2; z_2 - z_1, \mathbf{x}_1 - \frac{\mathbf{y}}{\theta}\Big)\, \Lambda(z_1, s, \mathbf{x}_1, \mathbf{y})\; ds\, d\mathbf{y}\,,$$

using the approximation in (3.21) for the filter and where we introduce

$$A_D = \sqrt{A^2 + \frac{2D}{3}}\,.$$

We choose $d = 2$, and in the nondimensionalized coordinates we let

$$f_0(t) = e^{-\frac{t^2\mathcal{D}}{2}} \cos(f_c t)\,, \tag{6.1}$$

with $\mathcal{D} = 5$, and moreover choose

$$z_1 = 1\,, \quad \mathbf{x}_c = \mathbf{x}_2 = \mathbf{0}\,, \quad f_c = 30\,, \quad \sigma_x = 1\,.$$

In Figure 6.1, we plot $\mathcal{I}$ in the case with a homogeneous medium. We use the parameter values $A = A_D = 20$ in the left plot. The estimation then captures the wavefield and wavefront for relatively large lateral offsets for the Green's function. The right plot corresponds to the situation with a small aperture, $A = A_D = 1.5$, in which case the wavefield and corresponding "moveout" are not captured.

In Figure 6.2 we plot $\mathcal{I}$ in the case with a random medium. We use the parameter values $A = 1.5, A_D = 9.3$. We then recapture a large part of the wavefield and wavefront (following a hyperbolic "moveout").

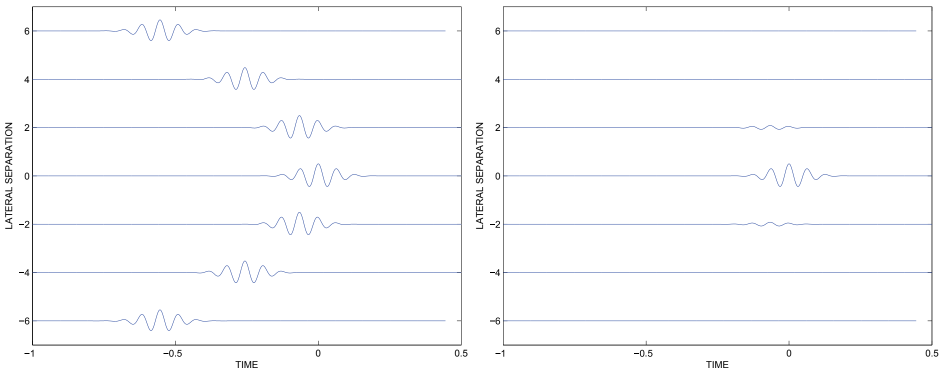


Fig. 6.1. *The normal moveout for a homogeneous medium and large (left) or small (right) aperture.*

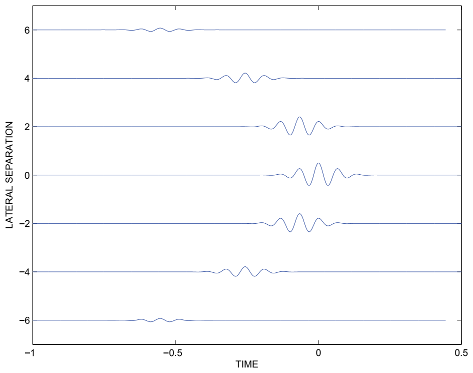


Fig. 6.2. *The normal moveout for a random medium with relatively large effective aperture.*

7. Conclusions. Pairwise cross-correlations between receivers forming arrays provide invaluable data sets where (deterministic) sources (earthquakes) are necessarily absent. The data sets are used to carry out tomography, or inverse scattering, to reveal the properties of the medium away from (below) the array. We tailored a scaling regime to applications in global earth seismology, generating "empirical" Green's functions and "virtual" source experiments. Characterization and knowledge about the structure of the incoherent waves and their correlations is crucial in this context. We have presented a first analysis of this approach to data acquisition in the framework of the paraxial wave approximation in a random medium. The analysis exploits the connection between cross-correlations and time-reversal. The connection is direct in the sense that the Green's function estimation problem can be articulated as the dual of time-reversal superresolution.

Important questions remain, for instance, how to design "optimal" filters, that is, how to combine optimally data in space and frequency to obtain stable and high-resolution estimates so that the filter Λ in (3.19) is near the identity, while enforcing statistical stability.

Appendix. Elements of the proof of Proposition 2.2. Consider the quantity

$$I_\delta(z,\mathbf{x},\mathbf{y}) = \iint W_\delta(z,\mathbf{x},\mathbf{p}) e^{-i\mathbf{p}\cdot\mathbf{y}}\, d\mathbf{p}\,.$$

The initial condition in (3.17),

$$W_\delta(0,\mathbf{x},\mathbf{p}) = W_I(\mathbf{x},\mathbf{p})\,,$$

is assumed to be uniformly bounded, Lipschitz continuous, and positive. As explained in [33] we can then write

$$W_\delta(z,\mathbf{x},\mathbf{p}) = W_I(\mathbf{X}_\delta(z,\mathbf{x},\mathbf{p}),\mathbf{P}_\delta(z,\mathbf{x},\mathbf{p}))\,,$$

for the stochastic flow $(\mathbf{X},\mathbf{P})$ satisfying

$$d\mathbf{X}_z = -\frac{1}{k}\mathbf{P}\,dz\,, \quad d\mathbf{P}_z = -\frac{k}{2}d\mathbf{B}(z)\,, \quad \mathbf{X}_0 = \mathbf{x}\,, \quad \mathbf{P}_0 = \mathbf{p}\,.$$

Observe first that

$$\iint \mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p})\right\} e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}\,d\mathbf{p} = \iint \mathbb{E}\left\{W_I(\mathbf{X}_\delta(z,\mathbf{x},\mathbf{p}),\mathbf{P}_\delta(z,\mathbf{x},\mathbf{p}))\right\} e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}\,d\mathbf{p}$$

is finite and independent of δ by (2.21). Writing the complex exponential in terms of its real and imaginary parts and decomposing the domain of integration for the corresponding integrals into subsets where $\cos(\mathbf{p}\cdot\mathbf{y})$ (respectively, $\sin(\mathbf{p}\cdot\mathbf{y})$) is positive (respectively, negative), we can apply Tonelli's theorem and interchange the order of integration and expectation and get

$$\mathbb{E}\{I_\delta(z,\mathbf{x},\mathbf{y})\} = \iint \mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p})\right\} e^{-\mathrm{i}\mathbf{p}\cdot\mathbf{y}}\,d\mathbf{p}; \tag{A.1}$$

thus, $I_\delta(z,\mathbf{x},\mathbf{y})$ is finite with probability one, and its expectation given by (A.1).

By a corresponding application of Tonelli's theorem we can write

$$\mathbb{E}\{I_\delta^2(z,\mathbf{x},\mathbf{y})\} = \iint \mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\right\} e^{-\mathrm{i}(\mathbf{p}_1+\mathbf{p}_2)\cdot\mathbf{y}}\,d\mathbf{p}_1\,d\mathbf{p}_2\,.$$

In [33] it is shown that $\mathbb{E}\{W_\delta^2(z,\mathbf{x},\mathbf{p})\}$ is integrable in $\mathbf{p}$, and that

$$\lim_{\delta\to 0}\mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p}_1)W_\delta(z,\mathbf{x},\mathbf{p}_2)\right\} = \mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p}_1)\right\}\mathbb{E}\left\{W_\delta(z,\mathbf{x},\mathbf{p}_2)\right\}.$$

By the Lebesgue dominated convergence theorem we can therefore conclude that

$$\lim_{\delta\to 0}\mathbb{E}\{I_\delta^2(z,\mathbf{x},\mathbf{y})\} = \mathbb{E}^2\{I_\delta(z,\mathbf{x},\mathbf{y})\}\,.$$

Acknowledgment. The authors would like to thank H. Yao for composing Figure 1.1.

REFERENCES

[1] D. V. Alfaro and K. Solna, *Time-Reversal for Inclusion Detection in Randomly Layered Media*, submitted, 2008.

[2] G. Bal, G. Papanicolaou, and L. Ryzhik, *Radiative transport limit for the Schrödinger equation*, Nonlinearity, 15 (2002), pp. 513–529.

[3] C. Bardos, J. Garnier, and G. Papanicoalou, *Identification of Green's function singularities by cross correlation of noisy signals*, Inverse Problems, 24 (2008), paper 015011.

[4] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Amer., 111 (2002), pp. 230–248.

[5] L. Borcea, G. Papanicolaou, and C. Tsogka, *Asymptotics for the space-time Wigner transform with applications to imaging*, in Stochastic Differential Equations: Theory and Applications, P. H. Baxendale and S. V. Lototsky, eds., Interdiscip. Math. Sci. 2, World Scientific, River Edge, NJ, 2007, pp. 91–112.

[6] L. Borcea, G. Papanicolaou, and C. Tsogka, *Coherent interferometric imaging*, Geophysics, 71 (2006), pp. S1165–S1175.

[7] M. Campillo and A. Paul, *Long-range correlations in the diffuse seismic coda*, Science, 299 (2003), pp. 547–549.

[8] J. F. Claerbout, *Synthesis of a layered medium from its acoustic transmission response*, Geophysics, 33 (1968), pp. 264–269.

[9] J. F. Claerbout, *Coarse grid calculations of waves in inhomogeneous media with application to delineation of complicated seismic structure*, Geophysics, 35 (1970), pp. 407–418.

[10] Y. Colin de Verdière, *Mathematical Models for Passive Imaging* I: *General Background*, preprint, 2006; available online at http://www-fourier.ujf-grenoble.fr/ycolver/Sismo/big1.pdf.

[11] D. Dawson and G. Papanicolaou, *A random wave process*, Appl. Math. Optim., 12 (1984), pp. 97–114.

[12] M. V. de Hoop and A. T. de Hoop, *Wave-field reciprocity and optimization in remote sensing*, Proc. R. Soc. Lond. A., 456 (2000), pp. 641–682.

[13] A. Derode and M. Fink, *How to estimate the Green's function of a heterogeneous medium between two passive sensors? Application to acoustic waves*, Appl. Phys. Lett., 83 (2003), pp. 3054–3056.

[14] A. Derode, E. Larose, M. Tanter, J. de Rosny, A. Tourin, M. Campillo, and M. Fink, *Recovering the Green's function from field-field correlations in an open scattering medium*, J. Acoust. Soc. Amer., 113 (2003), pp. 2973–2976.

[15] D. Draganov, K. Wapenaar, W. Mulder, J. Singer, and A. Verdel, *Retrieval of reflections from seismic background-noise measurements*, Geophys. Res. Lett., 34 (2007), paper L04305; doi:10.1029/2006GL028735.

[16] T. L. Duvall, S. M. Jefferies, J. W. Harvey, and A. Pomerantz, *Time distance helioseismology*, Nature, 362 (1993), pp. 430–432.

[17] A. Fannjiang, *White-noise and geometrical optics limits of Wigner–Moyal equation for beam waves in turbulent media.* II. *Two-frequency formulation*, J. Statist. Phys., 120 (2005), pp. 543–586.

[18] A. Fannjiang, *Two-frequency radiative transfer and asymptotic solution*, J. Opt. Soc. Amer. A, 24 (2007), pp. 2248–2256.

[19] A. Fannjiang and K. Solna, *Scaling limits for laser beam propagation in atmospheric turbulence*, Stoch. Dyn., 4 (2004), pp. 135–150.

[20] A. C. Fannjiang and K. Solna, *Propagation and time reversal of wave beams in atmospheric turbulence*, Multiscale Model. Simul., 3 (2005), pp. 522–558.

[21] A. Fannjiang and K. Solna, *Superresolution and duality for time-reversal of waves in random media*, Phys. Lett. A, 352 (2005), pp. 22–29.

[22] M. Fink, *Time-reversed acoustics*, Scientific American, Nov. (1999), pp. 91–97.

[23] J. P. Fouque, J. Garnier, A. Nachbin, and K. Sølna, *Imaging of a dissipative layer in a random medium using a time reversal method*, in Proceedings of the conference "Monte Carlo and Quasi-Monte Carlo Methods 2004," Nice, 2004, H. Niederreiter and D. Talay, eds., Springer, Berlin, 2005, pp. 127–145.

[24] J. P. Fouque, J. Garnier, A. Nachbin, and K. Solna, *Time reversal refocusing for point source in randomly layered media*, Wave Motion, 42 (2005), pp. 191–288.

[25] J. P. Fouque, J. Garnier, G. Papanicolaou, and K. Solna, *Wave Propagation and Time Reversal in Randomly Layered Media*, Springer, New York, 2007.

[26] J. Garnier, *Imaging in randomly layered media by cross-correlating noisy signals*, Multiscale Model. Simul., 4 (2005), pp. 610–640.

[27] P. Gerstoft, K. G. Sabra, P. Roux, W. A. Kuperman, and M. C. Fehler, *Green's function extraction and surface-wave tomography from microseisms in southern California*, Geophysics, 71 (2006), pp. S123–S131.

[28] R. He, B. E. Hornby, and G. Schuster, 3*D wave-equation interferometric migration of VSP multiples*, SEG Expanded Abstracts, 25 (2006), 3442.

[29] K. Huang, G. Papanicolaou, K. Sølna, C. Tsogka, and H. Zhao, *Efficient numerical simulation for long range wave propagation*, J. Comput. Phys., 215 (2006), pp. 448–464.

[30] W. A. Kuperman, W. S. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, *Phase conjugation in the ocean: Experimental demonstration of an acoustic time-reversal mirror*, J. Acoust. Soc. Amer., 103 (1997), pp. 25–40.

[31] G. Lerosey, J. de Rosny, A. Tourin, A. Derode, and M. Fink, *Time reversal of wideband microwaves*, Appl. Phys. Lett., 88 (2006), paper 154101.

[32] M. Levy, *Parabolic Equation Methods for Electromagnetic Wave Propagation*, The Institution of Electrical Engineers, Herts, UK, 2000.

[33] G. Papanicolaou, L. Ryzhik, and K. Sølna, *Statistical stability in time reversal*, SIAM J. Appl. Math., 64 (2004), pp. 1133–1155.

[34] G. Papanicolaou, L. Ryzhik, and K. Sølna, *Self-averaging from lateral diversity in the Itô–Schrödinger equation*, Multiscale Model. Simul., 6 (2007), pp. 468–492.

[35] P. Roux, K. G. Sabra, P. Gerstoft, W. A. Kuperman, and M. Fehler, *P-waves from cross correlation of seismic noise*, Geophys. Res. Lett., 32 (2005), paper L19303; doi:10.1029/2005GL023803.

[36] K. Sabra, P. Roux, and W. A. Kuperman, *Arrival-time structure of the time averaged ambient noise cross-correlation function in an ocean waveguide*, J. Acoust. Soc. Amer., 117 (2005), pp. 164–174.

[37] F. Scherbaum, *Seismic imaging of the site response using micro-earthquake recordings. Part* II: *Application to the Swabian Jura, Southwest Germany, seismic network*, Bull. Seismol. Soc. Am., 77 (1987), pp. 1924–1944.

[38] N. M. Shapiro and M. Campillo, *Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise*, Geophys. Res. Lett., 31 (2004), paper L07614.

[39] N. M. Shapiro, M. Campillo, L. Stehly, and M. Ritzwoller, *High-resolution surface-wave tomography from ambient seismic noise*, Science, 307 (2005), pp. 1615–1618.

[40] K. Sølna and G. C. Papanicolaou, *Ray theory for a locally layered medium*, Waves in Random Media, 10 (2000), pp. 151–198.

[41] F. D. Tappert, *The parabolic approximation method*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Phys. 70, Springer-Verlag, Berlin, 1977.

[42] B. A. van Tiggelen, *Green function retrieval and time reversal in a disordered world*, Phys. Rev. Lett., 91 (2003), paper 243904.

[43] H. Yao, R. D. van der Hilst, and M. V. de Hoop, *Surface-wave array tomography in SE Tibet from ambient seismic noise and two-station analysis*—I. *Phase velocity maps*, Geophys. J. Int., 166 (2006), pp. 732–744.

[44] J. Yu and G. T. Schuster, *Crosscorrelogram migration of inverse vertical seismic profile data*, Geophysics, 71 (2006), pp. S1–S11.

[45] K. Wapenaar, J. Fokkema, and R. Snieder, *Retrieving the Green's function in an open system by cross correlation: A comparison of approaches (L)*, J. Acoust. Soc. Amer., 118 (2005), pp. 2783–2786.

[46] R. L. Weaver and O. I. Lobkis, *Ultrasonics without a source: Thermal fluctuation correlations at Mhz frequencies*, Phys. Rev. Lett., 93 (2001), paper 254301.

[47] M. E. Willis, R. Lu, X. Campman, M. N. Toksöz, Y. Zhang, and M. V. de Hoop, *A novel application of time-reversed acoustics: Salt-dome flank imaging using walkaway VSP surveys*, Geophysics, 71 (2006), A7–A11.

SIAM J. APPL. MATH.
Vol. 69, No. 4, pp. 933–958

ON STABLE, COMPLETE, AND SINGULARITY-FREE BOUNDARY INTEGRAL FORMULATIONS OF EXTERIOR STOKES FLOW*

O. GONZALEZ†

Abstract. A new boundary integral formulation of the second kind for exterior Stokes flow is introduced. The formulation is stable, complete, singularity-free, and natural for bodies of complicated shape and topology. We prove an existence and uniqueness result for the formulation for arbitrary flows and illustrate its performance via several numerical examples using a Nyström method with Gauss–Legendre quadrature rules of different order.

Key words. Stokes equations, boundary integral equations, single-layer potentials, double-layer potentials, parallel surfaces, Nyström discretization

AMS subject classifications. 31B10, 35Q30, 76D07, 65R20

DOI. 10.1137/070698154

1. Introduction.

In this article we study boundary integral formulations of exterior Stokes flow problems around arbitrary bodies with prescribed velocity data. For such problems it is well known that a formulation based on either of the classic single- or double-layer Stokes potentials is inadequate [29, 30]. A formulation based on the single-layer potential leads to a boundary integral operator which is unstable in the sense that its condition number is unbounded and incomplete in the sense that its range is deficient. Consequently, such a formulation is not optimal for numerical discretization and not capable of representing an arbitrary exterior flow. A formulation based on the double-layer potential leads to a boundary integral operator which is stable in the sense that its condition number is bounded but which is incomplete—even more so than the single-layer potential. Thus, in contrast to the single-layer case, a double-layer formulation is optimal for numerical discretization, but like the single-layer case, it is not capable of representing an arbitrary exterior flow.

Various authors have shown that a double-layer formulation can be modified so as to obtain completeness while retaining stability [14, 18, 20, 26, 28]. In Power and Miranda [28] it was shown that a complete formulation can be obtained by adding two classic singular flow solutions (a stokeslet and rotlet) to the double-layer potential, where the poles of the singular solutions are coincident and placed at an arbitrary location within the body. In Hebeker [14] it was shown that a complete formulation can be obtained by simply taking a positive linear combination of the classic single- and double-layer potentials. The approach of Power and Miranda has the desirable feature of being singularity-free in the sense that it leads to an integral equation involving only bounded integrands. In contrast, the approach of Hebeker leads to an integral equation with unbounded integrands. On the other hand, the approach of Power and Miranda is not natural for flows around bodies of complex shape or topology for which there is no distinguished point for the stokeslet and rotlet pole. In contrast, the approach of Hebeker is natural for flows around bodies of arbitrary shape.

*Received by the editors July 23, 2007; accepted for publication (in revised form) August 21, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/siap/69-4/69815.html

†Department of Mathematics, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 (og@math.utexas.edu).

Here we introduce a new boundary integral formulation for exterior Stokes flow which combines the strengths of the Power and Miranda and the Hebeker formulations. The new formulation is stable, complete, singularity-free, and natural for bodies of complicated shape and topology. The formulation is made complete by virtue of a positive linear combination of single- and double-layer potentials and is made singularity-free by mapping the single-layer potential onto an appropriate parallel surface. We prove an existence and uniqueness result for the formulation for arbitrary flows and illustrate its performance via several numerical examples using a standard Nyström method based on Gauss–Legendre quadrature rules. Our results show that a standard method applied to the singularity-free formulation provides a simple and viable alternative to specialized methods required by classic formulations.

Classic boundary integral formulations of the Stokes equations involve weakly singular kernels that require special treatment. Such formulations can be treated with variants of the Nyström method which employ kernel-adapted product integration rules [3, 19] or coordinate transformations and projections which effectively remove the singularity [33]. Several types of Galerkin and collocation methods [3, 5, 6, 19] can also be applied to these formulations, as well as spectral Galerkin [2, 10, 12] and wavelet-based methods [1, 21, 32]. However, these approaches generally require basis functions that may be difficult to construct or which may exist only for certain classes of geometries. Moreover, they require special techniques for computing weakly singular integrals, which can be expensive. Here we show that such issues associated with classic formulations can be avoided in a simple and efficient way by a straightforward discretization of the singularity-free formulation.

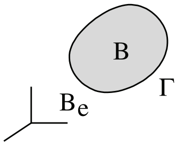
The presentation is structured as follows. In section 2 we outline the Stokes equations for the steady flow of an incompressible viscous fluid in an exterior domain. In sections 3 and 4 we establish notation and collect several results on singular solutions and surface potentials for the Stokes equations that will be employed throughout our developments. In sections 5 and 6 we summarize, for purposes of comparison, the Hebeker and the Power and Miranda formulations of the exterior Stokes problem and highlight several of their properties. In section 7 we introduce our new formulation and establish its solvability properties for arbitrary data. In section 8 we describe a numerical discretization of our formulation using a standard Nyström method with an arbitrary quadrature rule. In section 9 we illustrate our approach with numerical examples and summarize our conclusions.

2. The exterior Stokes problem.

In this section we define the boundary-value problem that we will study. We briefly outline standard assumptions which guarantee existence and uniqueness of solutions, and we introduce various flow quantities of interest that will be used to understand the properties of different boundary integral formulations.

2.1. Problem formulation.

We consider the steady motion of a body of arbitrary shape through an incompressible viscous fluid at a low Reynolds number. In a body-fixed frame, we denote the body domain by B, the fluid domain exterior to the body by B_e, and the body-fluid interface by Γ. Given a body velocity field $v : \Gamma \to \mathbb{R}^3$, the basic problem is to find a fluid velocity field $u : B_e \to \mathbb{R}^3$ and pressure field $p : B_e \to \mathbb{R}$ which satisfy the classic Stokes equations, which in nondimensional form are



$$\begin{aligned} &u_{i,jj} - p_{,i} = 0, && x \in B_e, \\ &u_{i,i} = 0, && x \in B_e, \\ &u_i = v_i, && x \in \mathit{\Gamma}, \\ &u_i, p \to 0, && |x| \to \infty. \end{aligned} \tag{2.1}$$

Equation $(2.1)_1$ is the local balance law of linear momentum for the fluid and $(2.1)_2$ is the local incompressibility constraint. Equation $(2.1)_3$ is the no-slip boundary condition which states that the fluid and body velocities coincide at each point of the boundary. The limits in $(2.1)_4$ are boundary conditions which are consistent with the fluid being at rest at infinity. Unless mentioned otherwise, all vector quantities are referred to a single basis and indices take values from one to three. Here and throughout we will use the usual conventions that a pair of repeated indices implies summation and that indices appearing after a comma denote partial derivatives.

2.2. Solvability. We assume $B \cup \mathit{\Gamma} \cup B_e$ fills all of three-dimensional space, B is open and bounded, and B_e is open and connected. Moreover, we assume $\mathit{\Gamma}$ consists of a finite number of disjoint, closed, bounded, and orientable components, each of which is a Lyapunov surface [13]. These conditions on $\mathit{\Gamma}$ imply that standard results from potential theory for the Stokes equations may be applied [20, 26, 29]. Moreover, together with the continuity of v, they are sufficient to guarantee that (2.1) has a unique solution (u, p) with the following decay properties [9, 20]:

$$u_i = O(|x|^{-1}), \quad u_{i,j} = O(|x|^{-2}), \quad p = O(|x|^{-2}) \quad \text{as} \quad |x| \to \infty. \tag{2.2}$$

The solution (u, p) is smooth in B_e but may possess only a finite number of bounded derivatives in $B_e \cup \mathit{\Gamma}$ depending on the precise smoothness of $\mathit{\Gamma}$ and v.

2.3. Basic flow quantities. The volume flow rate associated with a flow (u, p) and a given oriented surface S is defined by

$$Q = \int_S u_i(x) n_i(x)\, dA_x, \tag{2.3}$$

where $n : S \to \mathbb{R}^3$ is a given unit normal field and dA_x denotes an infinitesimal area element at $x \in S$. When S is closed and bounded, we always choose n to be the outward unit normal. In this case, Q quantifies the volume expansion rate of the domain enclosed by S.

The fluid stress field associated with a flow (u, p) is a function $\sigma : B_e \to \mathbb{R}^{3\times 3}$ defined by

$$\sigma_{ij} = -p\delta_{ij} + u_{i,j} + u_{j,i}, \tag{2.4}$$

where δ_{ij} is the standard Kronecker delta symbol. For each $x \in B_e$ the stress tensor σ is symmetric in the sense that $\sigma_{ij} = \sigma_{ji}$. The traction field $f : S \to \mathbb{R}^3$ exerted by the fluid on a given oriented surface S is defined by

$$f_i = \sigma_{ij} n_j. \tag{2.5}$$

The resultant force F and torque T, about an arbitrary point c, associated with f are

$$F_i = \int_S f_i(x)\, dA_x, \qquad T_i = \int_S \varepsilon_{ijk}(x_j - c_j) f_k(x)\, dA_x, \tag{2.6}$$

where ε_{ijk} is the standard permutation symbol. As before, when S is closed and bounded, we always choose n to be the outward unit normal field. In this case, F and T are loads exerted on S by the fluid exterior to S.

For convenience, we assume all quantities have been nondimensionalized using a characteristic length scale $\ell > 0$, a velocity scale $\vartheta > 0$, and a force scale $\mu\vartheta\ell > 0$, where μ is the absolute viscosity of the fluid. The dimensional quantities corresponding to $\{x, u, p, v\}$ are $\{\ell x, \vartheta u, \mu\vartheta\ell^{-1}p, \vartheta v\}$, and the dimensional quantities corresponding to $\{Q, \sigma, f, F, T\}$ are $\{\vartheta\ell^2 Q, \mu\vartheta\ell^{-1}\sigma, \mu\vartheta\ell^{-1}f, \mu\vartheta\ell F, \mu\vartheta\ell^2 T\}$.

3. Singular solutions of the Stokes equations. In this section we outline various classic singular solutions of the homogeneous, free-space Stokes equations

$$\begin{aligned} &u_{i,jj} - p_{,i} = 0, && x \neq y,\\ &u_{i,i} = 0, && x \neq y,\\ &u_i, p \to 0, && |x| \to \infty. \end{aligned} \tag{3.1}$$

Here y is a given point called the pole of the solution. Various representations of the solution of (2.1) can be derived and understood in terms of these solutions and their properties. Notice that, by linearity, any multiple or linear combination of solutions of (3.1) is also a solution where defined. In what follows, we let $z = x - y$ and $r = |z|$, and we let S_{int} and S_{ext} denote the interior and exterior domains associated with a given closed, bounded surface S. The notation and results outlined here will be employed throughout our developments.

3.1. Point-source solution. The point-source solution is defined by $u_i = U_i^{\text{PS}}$, $p = \varPi^{\text{PS}}$, $\sigma_{ik} = \varXi_{ik}^{\text{PS}}$, where

$$U_i^{\text{PS}} = \frac{z_i}{r^3}, \qquad \varPi^{\text{PS}} = 0, \qquad \varXi_{ik}^{\text{PS}} = \frac{2\delta_{ik}}{r^3} - \frac{6z_iz_k}{r^5}. \tag{3.2}$$

This solution may be derived from (3.1) by making the ansatz $u_i = \phi_{,i}$ and $p = 0$ for some radially symmetric function ϕ [30]. The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with an arbitrary closed, bounded surface S can be found by direct computation and depend on the relative location of the pole y. When $y \in S_{\text{ext}}$ the divergence theorem and (3.1) can be used to show that the relevant integrals over S all vanish. When $y \in S_{\text{int}}$ the divergence theorem and (3.1) can be used to transform the relevant integrals over S into integrals over an arbitrary sphere in S_{int} centered at y, which can then be evaluated directly. The results are

$$\begin{aligned} &F_i = 0, \quad && T_i = 0, \quad && Q = 4\pi, \quad && y \in S_{\text{int}},\\ &F_i = 0, && T_i = 0, && Q = 0, && y \in S_{\text{ext}}. \end{aligned} \tag{3.3}$$

3.2. Point-source dipole solution. The point-source dipole solution is defined by $u_i = U_{ij}^{\text{PSD}} g_j$, $p = \varPi_j^{\text{PSD}} g_j$, $\sigma_{ik} = \varXi_{ikj}^{\text{PSD}} g_j$, where g_j is an arbitrary vector independent of x and

$$\begin{aligned} &U_{ij}^{\text{PSD}} := \frac{\partial}{\partial y_j} U_i^{\text{PS}} = -\frac{\delta_{ij}}{r^3} + \frac{3z_iz_j}{r^5}, \qquad \varPi_j^{\text{PSD}} := \frac{\partial}{\partial y_j} \varPi^{\text{PS}} = 0,\\ &\varXi_{ikj}^{\text{PSD}} := \frac{\partial}{\partial y_j} \varXi_{ik}^{\text{PS}} = \frac{6(\delta_{ik}z_j + \delta_{ij}z_k + \delta_{kj}z_i)}{r^5} - \frac{30z_iz_kz_j}{r^7}. \end{aligned} \tag{3.4}$$

This solution is implied by the solution in (3.2) and the linearity of (3.1). The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with an arbitrary closed, bounded surface S can be computed as previously described. The results are

$$\begin{array}{llll} F_i = 0, & T_i = 0, & Q = 0, & y \in S_{\text{int}}, \\ F_i = 0, & T_i = 0, & Q = 0, & y \in S_{\text{ext}}. \end{array} \tag{3.5}$$

3.3. Point-force solution: Stokeslet. The point-force solution is defined by $u_i = U_{ij}^{\text{PF}} g_j$, $p = \Pi_j^{\text{PF}} g_j$, $\sigma_{ik} = \Xi_{ikj}^{\text{PF}} g_j$, where g_j is an arbitrary vector independent of x and

$$U_{ij}^{\text{PF}} = \frac{\delta_{ij}}{r} + \frac{z_i z_j}{r^3}, \qquad \Pi_j^{\text{PF}} = \frac{2z_j}{r^3}, \qquad \Xi_{ikj}^{\text{PF}} = -\frac{6 z_i z_k z_j}{r^5}. \tag{3.6}$$

Up to a normalizing constant, this solution corresponds to the classic fundamental solution of (3.1) and can be derived using the technique of Fourier transforms [20, 30]. It is typically referred to as a *stokeslet.* The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with an arbitrary closed, bounded surface S can be computed as previously described. The results are

$$\begin{array}{llll} F_i = -8\pi g_i, & T_i = -8\pi \varepsilon_{ijk}(y_j - c_j) g_k, & Q = 0, & y \in S_{\text{int}}, \\ F_i = 0, & T_i = 0, & Q = 0, & y \in S_{\text{ext}}. \end{array} \tag{3.7}$$

3.4. Point-force dipole solution: Stresslet, rotlet. The point-force dipole solution is defined by $u_i = U_{ijl}^{\text{PFD}} g_{jl}$, $p = \Pi_{jl}^{\text{PFD}} g_{jl}$, $\sigma_{ik} = \Xi_{ikjl}^{\text{PFD}} g_{jl}$, where g_{jl} is an arbitrary tensor independent of x and

$$\begin{aligned} U_{ijl}^{\text{PFD}} &:= \frac{\partial}{\partial y_l} U_{ij}^{\text{PF}} = \frac{\delta_{ij} z_l - \delta_{il} z_j - \delta_{jl} z_i}{r^3} + \frac{3 z_i z_j z_l}{r^5}, \\ \Pi_{jl}^{\text{PFD}} &:= \frac{\partial}{\partial y_l} \Pi_j^{\text{PF}} = -\frac{2\delta_{jl}}{r^3} + \frac{6 z_j z_l}{r^5}, \\ \Xi_{ikjl}^{\text{PFD}} &:= \frac{\partial}{\partial y_l} \Xi_{ikj}^{\text{PF}} = \frac{6(\delta_{il} z_k z_j + \delta_{kl} z_i z_j + \delta_{jl} z_i z_k)}{r^5} - \frac{30 z_i z_k z_j z_l}{r^7}. \end{aligned} \tag{3.8}$$

This solution is implied by the solution in (3.6) and the linearity of (3.1). By considering the decomposition $g_{jl} = g_{jl}^{sym} + g_{jl}^{skw}$, where $g_{jl}^{sym} = \frac{1}{2}(g_{jl} + g_{lj})$ and $g_{jl}^{skw} = \frac{1}{2}(g_{jl} - g_{lj})$, and by using the parameterization $g_{jl}^{skw} = \frac{1}{2}\varepsilon_{jml} g_m^{vec}$, we find that the point-force dipole solution can be decomposed as

$$\begin{aligned} U_{ijl}^{\text{PFD}} g_{jl} &= -U_i^{\text{PS}} \delta_{jl} g_{jl}^{sym} + U_{ijl}^{\text{STR}} g_{jl}^{sym} + U_{im}^{\text{ROT}} g_m^{vec}, \\ \Pi_{jl}^{\text{PFD}} g_{jl} &= -\Pi^{\text{PS}} \delta_{jl} g_{jl}^{sym} + \Pi_{jl}^{\text{STR}} g_{jl}^{sym} + \Pi_m^{\text{ROT}} g_m^{vec}. \end{aligned} \tag{3.9}$$

Here $(U_i^{\text{PS}}, \Pi^{\text{PS}})$ is the point-source solution given in (3.2) and $(U_{ijl}^{\text{STR}}, \Pi_{jl}^{\text{STR}})$ and $(U_{im}^{\text{ROT}}, \Pi_m^{\text{ROT}})$ are detailed below. By linearity, and the fact that g_{jl}^{sym} and g_m^{vec} are independent, we deduce that each of these pairs provides an independent solution of (3.1).

Stresslet solution. The stresslet solution is $u_i = U_{ijl}^{\text{STR}} h_{jl}$, $p = \Pi_{jl}^{\text{STR}} h_{jl}$, $\sigma_{ik} = \Xi_{ikjl}^{\text{STR}} h_{jl}$, where h_{jl} is an arbitrary tensor independent of x and

$$\begin{gathered} U_{ijl}^{\text{STR}} = \frac{3 z_i z_j z_l}{r^5}, \qquad \Pi_{jl}^{\text{STR}} = -\frac{2\delta_{jl}}{r^3} + \frac{6 z_j z_l}{r^5}, \\ \Xi_{ikjl}^{\text{STR}} = \frac{2\delta_{ik}\delta_{jl}}{r^3} + \frac{3(\delta_{ij} z_k z_l + \delta_{il} z_j z_k + \delta_{jk} z_i z_l + \delta_{lk} z_i z_j)}{r^5} - \frac{30 z_i z_j z_k z_l}{r^7}. \end{gathered} \tag{3.10}$$

Due to the symmetry of the above functions in the indices j and l we notice that only the symmetric part of h_{jl} contributes to the solution in concordance with (3.9). The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with an arbitrary closed, bounded surface S can be computed as previously described. The results are

$$\begin{array}{llll} F_i = 0, & T_i = 0, & Q = 4\pi h_{jj}, & y \in S_{\text{int}}, \\ F_i = 0, & T_i = 0, & Q = 0, & y \in S_{\text{ext}}. \end{array} \tag{3.11}$$

Rotlet (or couplet) solution. The rotlet solution is $u_i = U_{ij}^{\text{ROT}} h_j$, $p = \Pi_j^{\text{ROT}} h_j$, $\sigma_{ik} = \Xi_{ikj}^{\text{ROT}} h_j$, where h_j is an arbitrary vector independent of x and

$$U_{ij}^{\text{ROT}} = \frac{\varepsilon_{ijl} z_l}{r^3}, \qquad \Pi_j^{\text{ROT}} = 0, \qquad \Xi_{ikj}^{\text{ROT}} = \frac{3(\varepsilon_{ilj} z_k z_l + \varepsilon_{klj} z_i z_l)}{r^5}. \tag{3.12}$$

The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with an arbitrary closed, bounded surface S can be computed as previously described. The results are

$$\begin{array}{llll} F_i = 0, & T_i = -8\pi h_i, & Q = 0, & y \in S_{\text{int}}, \\ F_i = 0, & T_i = 0, & Q = 0, & y \in S_{\text{ext}}. \end{array} \tag{3.13}$$

Remarks 3.1.

1. It can be shown that all higher-order point-source solutions beginning with the dipole can be expressed in terms of the point-force solution [30]. In particular, we have

$$U_{ij}^{\text{PSD}} = -\frac{1}{2} \frac{\partial^2 U_{ij}^{\text{PF}}}{\partial y_k \partial y_k}, \qquad \Pi_j^{\text{PSD}} = -\frac{1}{2} \frac{\partial^2 \Pi_j^{\text{PF}}}{\partial y_k \partial y_k}.$$

   Thus the family of higher-order point-source solutions is contained within the family of higher-order point-force solutions.
2. One approach to solving the boundary-value problem in (2.1) is to consider a linear combination (discrete or continuous) of singular solutions with poles placed arbitrarily within the body domain B. The coefficients in the combination are then determined by enforcing the boundary condition on Γ. However, because arbitrary boundary conditions can in general not be satisfied exactly in this approach, it yields only approximate solutions of (2.1) [7, 30]. For example, slender-body theory is based on this approach [4, 15, 17].
3. A related approach to solving (2.1) is to consider a linear combination of singular solutions with poles distributed continuously over the surface Γ. The density of the distribution is then determined by enforcing the boundary condition on Γ. This approach leads to the classic theory of surface potentials for the Stokes equations and yields exact representations of the solutions of (2.1) [20, 26, 29, 30].

4. Surface potentials for the Stokes equations. In this section we outline the classic single- and double-layer surface potentials for the Stokes equations and summarize their main properties. All the boundary integral formulations that we will study are based on these potentials. In what follows Γ is an arbitrary closed, bounded surface with interior domain B and exterior domain B_e, as described in section 2.

4.1. Definition. Let $\psi : \mathit{\Gamma} \to \mathbb{R}^3$ be given. Then by the Stokes single-layer potentials on $\mathit{\Gamma}$ with density ψ we mean

$$\begin{aligned} V_i[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} U_{ij}^{\mathrm{PF}}(x,y)\psi_j(y)\,dA_y, \\ P_V[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} \mathit{\Pi}_j^{\mathrm{PF}}(x,y)\psi_j(y)\,dA_y, \end{aligned} \tag{4.1}$$

and by the Stokes double-layer potentials on $\mathit{\Gamma}$ with density ψ we mean

$$\begin{aligned} W_i[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} U_{ijl}^{\mathrm{STR}}(x,y)\psi_j(y)\nu_l(y)\,dA_y, \\ P_W[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} \mathit{\Pi}_{jl}^{\mathrm{STR}}(x,y)\psi_j(y)\nu_l(y)\,dA_y. \end{aligned} \tag{4.2}$$

Here $(U_{ij}^{\mathrm{PF}}, \mathit{\Pi}_j^{\mathrm{PF}})$ is the point-force or stokeslet solution in (3.6) with pole at y, $(U_{ijl}^{\mathrm{STR}}, \mathit{\Pi}_{jl}^{\mathrm{STR}})$ is the stresslet solution in (3.10) with pole at y, and ν is the unit normal field on $\mathit{\Gamma}$ directed outwardly from B. All densities ψ will be assumed continuous.

4.2. Analytic properties. For arbitrary density ψ the single-layer potentials $(V[\mathit{\Gamma},\psi], P_V[\mathit{\Gamma},\psi])$ and double-layer potentials $(W[\mathit{\Gamma},\psi], P_W[\mathit{\Gamma},\psi])$ are smooth at each $x \notin \mathit{\Gamma}$. Moreover, by virtue of their definitions as continuous linear combinations of stokeslets and stresslets, they satisfy the homogeneous Stokes equations $(2.1)_{1,2,4}$ at each $x \notin \mathit{\Gamma}$.

For arbitrary ψ the functions $V[\mathit{\Gamma},\psi]$ and $W[\mathit{\Gamma},\psi]$ are well defined for all $x \in B \cup \mathit{\Gamma} \cup B_e$. For $x \in \mathit{\Gamma}$ the integrands in $(4.1)_1$ and $(4.2)_1$ are unbounded functions of $y \in \mathit{\Gamma}$, but the integrals exist as improper integrals in the usual sense [13] provided that $\mathit{\Gamma}$ is a Lyapunov surface. The restrictions of $V[\psi,\mathit{\Gamma}]$ and $W[\psi,\mathit{\Gamma}]$ to $\mathit{\Gamma}$ are denoted by $\overline{V}[\psi,\mathit{\Gamma}]$ and $\overline{W}[\psi,\mathit{\Gamma}]$. These restrictions are continuous functions on $\mathit{\Gamma}$ [20]. Moreover, for any $x_0 \in \mathit{\Gamma}$ the following limit relations hold [20, 29, 30]:

$$\lim_{\substack{x\to x_0\\ x\in B_e}} V[\mathit{\Gamma},\psi](x) = \overline{V}[\mathit{\Gamma},\psi](x_0), \tag{4.3}$$

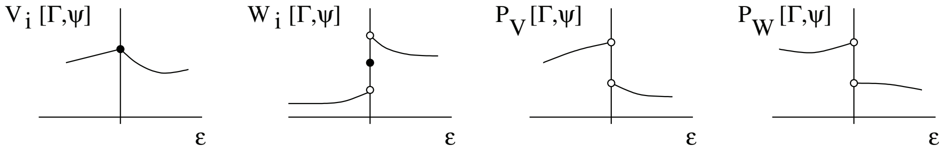
$$\lim_{\substack{x\to x_0\\ x\in B}} V[\mathit{\Gamma},\psi](x) = \overline{V}[\mathit{\Gamma},\psi](x_0), \tag{4.4}$$

$$\lim_{\substack{x\to x_0\\ x\in B_e}} W[\mathit{\Gamma},\psi](x) = \alpha\psi(x_0) + \overline{W}[\mathit{\Gamma},\psi](x_0), \tag{4.5}$$

$$\lim_{\substack{x\to x_0\\ x\in B}} W[\mathit{\Gamma},\psi](x) = -\alpha\psi(x_0) + \overline{W}[\mathit{\Gamma},\psi](x_0). \tag{4.6}$$

Here α is a constant that depends on the choice of normalization of the stresslet solution (3.10). For our choice we have $\alpha = 2\pi$. Notice that, by continuity of ψ and $\overline{W}[\mathit{\Gamma},\psi]$, the one-sided limits in (4.5) and (4.6) are themselves continuous functions on $\mathit{\Gamma}$.

In contrast to the case with $V[\mathit{\Gamma},\psi]$ and $W[\mathit{\Gamma},\psi]$, for arbitrary ψ the functions $P_V[\mathit{\Gamma},\psi]$ and $P_W[\mathit{\Gamma},\psi]$ do not exist as improper integrals in the usual sense when $x \in \mathit{\Gamma}$. In particular, the integrands in $(4.1)_2$ and $(4.2)_2$ are excessively singular functions of $y \in \mathit{\Gamma}$. Nevertheless, for sufficiently smooth $\mathit{\Gamma}$ and ψ, the functions $P_V[\mathit{\Gamma},\psi]$ and $P_W[\mathit{\Gamma},\psi]$ have well-defined limits as x approaches the surface $\mathit{\Gamma}$ [20, 29, 33]. Introducing $x_\epsilon = x_0 + \epsilon\nu(x_0)$, where $x_0 \in \mathit{\Gamma}$, the continuity properties of the functions $V[\mathit{\Gamma},\psi]$, $W[\mathit{\Gamma},\psi]$, $P_V[\mathit{\Gamma},\psi]$, $P_W[\mathit{\Gamma},\psi]$ around $\epsilon = 0$ can be illustrated as follows:



In general, the limits of the functions $V[\mathit{\Gamma},\psi]$, $W[\mathit{\Gamma},\psi]$, $P_V[\mathit{\Gamma},\psi]$, $P_W[\mathit{\Gamma},\psi]$ as x approaches $\mathit{\Gamma}$ from B_e or B have more physical significance than any directly defined values of these functions on $\mathit{\Gamma}$. In particular, physically meaningful boundary conditions are imposed on limit values and not on directly defined values. We remark that directly defined values of $P_V[\mathit{\Gamma},\psi]$ and $P_W[\mathit{\Gamma},\psi]$ on $\mathit{\Gamma}$ may be obtained by appealing to the theory of singular and hypersingular integrals [22, 24, 25].

4.3. Associated stress fields. For arbitrary ψ the stress fields associated with the single- and double-layer potentials are

$$\mathit{\Sigma}_V^{ik}[\mathit{\Gamma},\psi](x) = \int_{\mathit{\Gamma}} \mathit{\Xi}_{ikj}^{\mathrm{PF}}(x,y)\psi_j(y)\,dA_y, \tag{4.7}$$

$$\mathit{\Sigma}_W^{ik}[\mathit{\Gamma},\psi](x) = \int_{\mathit{\Gamma}} \mathit{\Xi}_{ikjl}^{\mathrm{STR}}(x,y)\psi_j(y)\nu_l(y)\,dA_y, \tag{4.8}$$

where $\mathit{\Xi}_{ikj}^{\mathrm{PF}}$ and $\mathit{\Xi}_{ikjl}^{\mathrm{STR}}$ are the stress functions corresponding to the point-force and stresslet solutions in (3.6) and (3.10). For arbitrary ψ the single-layer stress field $\mathit{\Sigma}_V[\mathit{\Gamma},\psi]$ is smooth at each $x \notin \mathit{\Gamma}$ and is the actual stress field associated with the Stokes flow with velocity field $V[\mathit{\Gamma},\psi]$ and pressure field $P_V[\mathit{\Gamma},\psi]$. A similar remark applies to the double-layer stress field $\mathit{\Sigma}_W[\mathit{\Gamma},\psi]$.

For $x \in \mathit{\Gamma}$ and arbitrary ψ the single-layer traction field $\mathit{\Sigma}_V[\mathit{\Gamma},\psi]\nu$ exists as an improper integral in the usual sense—but not the double-layer traction field $\mathit{\Sigma}_W[\mathit{\Gamma},\psi]\nu$. Moreover, for sufficiently smooth $\mathit{\Gamma}$ and ψ the following limit relations for $\mathit{\Sigma}_V[\mathit{\Gamma},\psi]\nu$ [20, 29] and $\mathit{\Sigma}_W[\mathit{\Gamma},\psi]\nu$ [29] hold for each $x_0 \in \mathit{\Gamma}$:

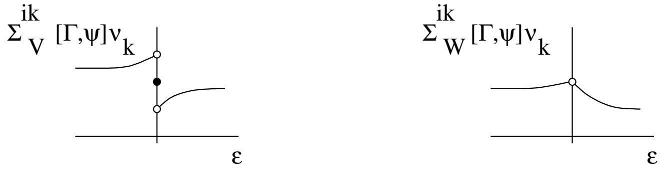
$$\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0) = \quad \beta\psi(x_0) + \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x_0)\nu(x_0), \tag{4.9}$$

$$\lim_{\substack{\epsilon\to 0\\ \epsilon<0}} \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0) = -\beta\psi(x_0) + \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x_0)\nu(x_0), \tag{4.10}$$

$$\lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \mathit{\Sigma}_W[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0) = \lim_{\substack{\epsilon\to 0\\ \epsilon<0}} \mathit{\Sigma}_W[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0). \tag{4.11}$$

Here $x_\epsilon = x_0 + \epsilon\nu(x_0)$ and β is a constant that depends on the choice of normalization of the point-force solution (3.6). For our choice we have $\beta = -4\pi$. The result in (4.11) is commonly referred to as the Lyapunov–Tauber condition.

For arbitrary ψ and $x_0 \in \mathit{\Gamma}$ the continuity properties of $\mathit{\Sigma}_V[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0)$ and $\mathit{\Sigma}_W[\mathit{\Gamma},\psi](x_\epsilon)\nu(x_0)$ around $\epsilon = 0$ can be illustrated as follows:



We remark that, as with $P_W[\mathit{\Gamma},\psi]$, a directly defined value of $\mathit{\Sigma}_W[\mathit{\Gamma},\psi]\nu$ on $\mathit{\Gamma}$ may be obtained by appealing to the theory of hypersingular integrals.

4.4. Flow properties.

4.4. Flow properties. Let S be an arbitrary closed, bounded surface with $\mathit{\Gamma} \subset S_{\mathrm{int}}$, and let n be the outward unit normal field on S. For arbitrary ψ the resultant force $F_V[\mathit{\Gamma},\psi]$, torque $T_V[\mathit{\Gamma},\psi]$ about an arbitrary point c, and volume flow rate $Q_V[\mathit{\Gamma},\psi]$ associated with S and the single-layer flow $(V[\mathit{\Gamma},\psi], P_V[\mathit{\Gamma},\psi])$ are

$$F_V[\mathit{\Gamma},\psi] = \int_S \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x)n(x)\,dA_x = -8\pi\int_{\mathit{\Gamma}} \psi(y)\,dA_y, \tag{4.12}$$

$$T_V[\mathit{\Gamma},\psi] = \int_S (x-c)\times \mathit{\Sigma}_V[\mathit{\Gamma},\psi](x)n(x)\,dA_x = -8\pi\int_{\mathit{\Gamma}} (y-c)\times\psi(y)\,dA_y, \tag{4.13}$$

$$Q_V[\mathit{\Gamma},\psi] = \int_S V[\mathit{\Gamma},\psi](x)\cdot n(x)\,dA_x = 0. \tag{4.14}$$

These results follow from the definitions of the single-layer stress and velocity fields in (4.7) and (4.1) and the properties of the point-force solution in (3.6) and (3.7) with g_i replaced by ψ_i. Because the above results are independent of S with $\mathit{\Gamma} \subset S_{\mathrm{int}}$, we can pass to the limit and conclude that the resultant force, torque, and volume flow rate associated with $\mathit{\Gamma}$ and the exterior single-layer flow are also given by the above results.

Similar calculations can be performed in the double-layer case. For arbitrary ψ the resultant force $F_W[\mathit{\Gamma},\psi]$, torque $T_W[\mathit{\Gamma},\psi]$ about an arbitrary point c, and volume flow rate $Q_W[\mathit{\Gamma},\psi]$ associated with S and the double-layer flow $(W[\mathit{\Gamma},\psi], P_W[\mathit{\Gamma},\psi])$ are

$$F_W[\mathit{\Gamma},\psi] = \int_S \mathit{\Sigma}_W[\mathit{\Gamma},\psi](x)n(x)\,dA_x = 0, \tag{4.15}$$

$$T_W[\mathit{\Gamma},\psi] = \int_S (x-c)\times \mathit{\Sigma}_W[\mathit{\Gamma},\psi](x)n(x)\,dA_x = 0, \tag{4.16}$$

$$Q_W[\mathit{\Gamma},\psi] = \int_S W[\mathit{\Gamma},\psi](x)\cdot n(x)\,dA_x = 4\pi\int_{\mathit{\Gamma}} \psi(y)\cdot\nu(y)\,dA_y. \tag{4.17}$$

These results follow from the definitions of the double-layer stress and velocity fields in (4.8) and (4.2) and the properties of the stresslet solution in (3.10) and (3.11) with h_{jl} replaced by $\psi_j\nu_l$. As before, because the above results are independent of S with $\mathit{\Gamma} \subset S_{\mathrm{int}}$, we can pass to the limit and conclude that the resultant force, torque, and volume flow rate associated with $\mathit{\Gamma}$ and the exterior double-layer flow are also given by the above results.

5. Hebeker formulation.

In this section we outline the boundary integral formulation of (2.1) introduced by Hebeker [14] and highlight several of its properties for comparison. In what follows $\mathit{\Gamma}$ is an arbitrary closed, bounded surface with interior domain B and exterior domain B_e, as described in section 2.

5.1. Formulation.

Given an arbitrary density $\psi : \mathit{\Gamma} \to \mathbb{R}^3$ and number $\theta \in [0,1]$, define $u : B_e \to \mathbb{R}^3$ and $p : B_e \to \mathbb{R}$ by

$$u = \theta V[\mathit{\Gamma},\psi] + (1-\theta)W[\mathit{\Gamma},\psi], \qquad p = \theta P_V[\mathit{\Gamma},\psi] + (1-\theta)P_W[\mathit{\Gamma},\psi]. \tag{5.1}$$

By properties of the single- and double-layer potentials, the fields (u,p) are smooth at each $x \in B_e$ and satisfy the Stokes equations $(2.1)_{1,2,4}$ at each $x \in B_e$. The stress field $\sigma : B_e \to \mathbb{R}^{3\times 3}$ associated with (u,p) is given by

$$\sigma = \theta \Sigma_V[\Gamma,\psi] + (1-\theta)\Sigma_W[\Gamma,\psi], \tag{5.2}$$

and the resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with Γ are

$$F = \theta F_V[\Gamma,\psi], \quad T = \theta T_V[\Gamma,\psi], \quad Q = (1-\theta)Q_W[\Gamma,\psi]. \tag{5.3}$$

Here we have used linearity and the flow properties of the single- and double-layer potentials outlined in section 4.4.

In order for (u,p) to provide the unique solution of the exterior Stokes boundary-value problem (2.1), the boundary condition $(2.1)_3$ must be satisfied. In particular, given $v : \Gamma \to \mathbb{R}^3$, we require

$$\lim_{\substack{x\to x_0\\ x\in B_e}} u(x) = v(x_0) \qquad \forall x_0 \in \Gamma. \tag{5.4}$$

Substituting for u from (5.1) and using the limit relations in (4.3) and (4.5), we obtain a boundary integral equation for the unknown density ψ:

$$\theta\overline{V}[\Gamma,\psi](x_0) + (1-\theta)\overline{W}[\Gamma,\psi](x_0) + (1-\theta)\alpha\psi(x_0) = v(x_0) \qquad \forall x_0 \in \Gamma. \tag{5.5}$$

From this we can deduce that (u,p) defined in (5.1) will be the unique solution of (2.1) if and only if ψ satisfies (5.5). This equation can be written in the standard form

$$\int_\Gamma K_\theta(x,y)\psi(y)\,dA_y + c_\theta\psi(x) = v(x) \qquad \forall x \in \Gamma, \tag{5.6}$$

where x_0 has been replaced by x for convenience, $c_\theta = (1-\theta)\alpha$, and

$$K_\theta^{ij}(x,y) = \theta U_{ij}^{\mathrm{PF}}(x,y) + (1-\theta)U_{ijl}^{\mathrm{STR}}(x,y)\nu_l(y). \tag{5.7}$$

Remarks 5.1.

1. Assuming Γ is a Lyapunov surface the kernel function $K_\theta(x,y)$ can be shown to be weakly singular. Thus the solvability of the linear integral equation (5.6) can be assessed via the Fredholm theory [19, 23]. Notice that $c_\theta = 0$ when $\theta = 1$, and $c_\theta \neq 0$ when $\theta \in [0,1)$. Thus (5.6) is a Fredholm equation of the first kind when $\theta = 1$ and of the second kind when $\theta \in [0,1)$.
2. The case $\theta = 0$ in (5.1) corresponds to a classic double-layer representation of (u,p). It is well known that this representation is incomplete in the sense that it can represent only those flows for which the resultant force and torque on Γ vanish, that is, $F = 0$ and $T = 0$ [20, 26, 29, 30]. Equivalently, the range of the linear operator in (5.6) is deficient, leading to solvability conditions and nonuniqueness for ψ.
3. The case $\theta = 1$ in (5.1) corresponds to a classic single-layer representation of (u,p). It is well known that this representation is also incomplete in the sense that it can represent only those flows for which the volumetric expansion rate of Γ vanishes, that is, $Q = 0$ [20, 26, 29, 30]. Equivalently, the range of the linear operator in (5.6) is again deficient, leading to solvability conditions and nonuniqueness for ψ.

4. The main idea in Hebeker [14] was to consider a mixed representation corresponding to $\theta \in (0,1)$. The intuitive motivation is that, by considering a linear combination, each potential can make up for the deficiencies of the other. As outlined below, such a representation is complete in the sense that it can represent arbitrary flows and stable in the sense that the density ψ depends continuously on the boundary data v.

5.2. Solvability result. The following is a slight generalization of the solvability result given in Hebeker [14].

Theorem 5.1 (see [14]). *Assume Γ is a closed, bounded Lyapunov surface. If $\theta \in (0,1)$, then* (5.6) *possesses a unique continuous solution ψ for any continuous boundary data v.*

Thus arbitrary solutions of the exterior Stokes boundary-value problem (2.1) can be represented in the form (5.1) with a unique density ψ for each $\theta \in (0,1)$. The presence of the double-layer potential in (5.1) ensures that the representation is stable. In particular, because (5.6) is a uniquely solvable Fredholm equation of the second kind, the linear operator in (5.6) has a finite condition number and the density ψ depends continuously on the data v. The presence of the single-layer potential in (5.1) ensures that the representation is complete. In particular, the single-layer potential completes the deficient range associated with the double-layer potential. The smoothness properties of the density ψ depend on those of the surface Γ and the data v.

Remarks 5.2.

1. Aside from the restriction of solvability, the parameter θ is arbitrary and can be exploited. For example, $\theta \in (0,1)$ might be chosen by some means to optimize the conditioning of the linear operator in (5.6).
2. Numerical methods for (5.6), Nyström methods in particular, must deal with the singularities in the kernels of the single- and double-layer potentials. The singularity in the kernel of the double-layer potential can be removed in a simple, standard way by employing a well-known integral identity [11, 28, 30, 31] (see section 8). However, there seems to be no similar removal technique for the singularity in the kernel of the single-layer potential.
3. In general numerical treatments, the singularity in the kernel of the single-layer potential can be dealt with by employing a kernel-adapted product quadrature rule [3, 19], together with a local coordinate transformation such as a Duffy transformation [3, 8, 30], or a floating polar transformation [33]. The same techniques can also be applied to the double-layer potential. In view of the inconvenience associated with the single-layer potential, we investigate alternative formulations.

6. Power and Miranda formulation.

In this section we outline the boundary integral formulation of (2.1) introduced by Power and Miranda [28] and highlight several of its properties for comparison. In what follows Γ is an arbitrary closed, bounded surface with interior domain B and exterior domain B_e, as described in section 2. For simplicity, in this section we assume that B consists of only one connected component. All the results outlined generalize in a straightforward way to the case when B has a finite number of disjoint components [29].

6.1. Formulation. Given an arbitrary density $\psi : \Gamma \to \mathbb{R}^3$, number $\theta \in [0,1]$, and point $x_* \in B$, define $u : B_e \to \mathbb{R}^3$ and $p : B_e \to \mathbb{R}$ by

$$u = \theta Y[\Gamma,\psi] + (1-\theta)W[\Gamma,\psi], \qquad p = \theta P_Y[\Gamma,\psi] + (1-\theta)P_W[\Gamma,\psi], \tag{6.1}$$

where $Y[\mathit{\Gamma},\psi]$ and $P_Y[\mathit{\Gamma},\psi]$ are fields defined in terms of the point-force (stokeslet) and rotlet solutions as

$$\begin{aligned} Y_i[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} \Big(U^{\mathrm{PF}}_{ij}(x,x_*) + U^{\mathrm{ROT}}_{il}(x,x_*)\varepsilon_{lpj}(y_p - x_{*p})\Big)\psi_j(y)\,dA_y, \\ P_Y[\mathit{\Gamma},\psi](x) &= \int_{\mathit{\Gamma}} \Big(\mathit{\Pi}^{\mathrm{PF}}_{j}(x,x_*) + \mathit{\Pi}^{\mathrm{ROT}}_{l}(x,x_*)\varepsilon_{lpj}(y_p - x_{*p})\Big)\psi_j(y)\,dA_y. \end{aligned} \tag{6.2}$$

By properties of the point-force and rotlet solutions and the double-layer potentials, the fields (u,p) are smooth at each $x \in B_e$ and satisfy the Stokes equations $(2.1)_{1,2,4}$ at each $x \in B_e$. The stress field $\sigma : B_e \to \mathbb{R}^{3\times 3}$ associated with (u,p) is given by

$$\sigma = \theta\mathit{\Sigma}_Y[\mathit{\Gamma},\psi] + (1-\theta)\mathit{\Sigma}_W[\mathit{\Gamma},\psi], \tag{6.3}$$

where $\mathit{\Sigma}_Y[\mathit{\Gamma},\psi]$ is the stress field associated with the flow $(Y[\mathit{\Gamma},\psi], P_Y[\mathit{\Gamma},\psi])$, namely,

$$\mathit{\Sigma}^{ik}_Y[\mathit{\Gamma},\psi](x) = \int_{\mathit{\Gamma}} \Big(\mathit{\Xi}^{\mathrm{PF}}_{ikj}(x,x_*) + \mathit{\Xi}^{\mathrm{ROT}}_{ikl}(x,x_*)\varepsilon_{lpj}(y_p - x_{*p})\Big)\psi_j(y)\,dA_y. \tag{6.4}$$

The resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with $\mathit{\Gamma}$ are

$$F = \theta F_Y[\mathit{\Gamma},\psi], \quad T = \theta T_Y[\mathit{\Gamma},\psi], \quad Q = \theta Q_Y[\mathit{\Gamma},\psi] + (1-\theta)Q_W[\mathit{\Gamma},\psi], \tag{6.5}$$

where $F_Y[\mathit{\Gamma},\psi]$, $T_Y[\mathit{\Gamma},\psi]$, and $Q_Y[\mathit{\Gamma},\psi]$ are the resultant force, torque, and volume flow rate associated with the flow $(Y[\mathit{\Gamma},\psi], P_Y[\mathit{\Gamma},\psi])$. From the properties of the point-force and rotlet solutions given in (3.7) and (3.13), we deduce that $Q_Y[\mathit{\Gamma},\psi] = 0$ and

$$F_Y[\mathit{\Gamma},\psi] = -8\pi\int_{\mathit{\Gamma}} \psi(y)\,dA_y, \quad T_Y[\mathit{\Gamma},\psi] = -8\pi\int_{\mathit{\Gamma}} (y-c)\times\psi(y)\,dA_y. \tag{6.6}$$

In order for (u,p) to provide the unique solution of the exterior Stokes boundary-value problem (2.1), the boundary condition $(2.1)_3$ must be satisfied. In particular, given $v : \mathit{\Gamma} \to \mathbb{R}^3$, we require

$$\lim_{\substack{x\to x_0\\ x\in B_e}} u(x) = v(x_0) \qquad \forall x_0 \in \mathit{\Gamma}. \tag{6.7}$$

Substituting for u from (6.1) and using the limit relation in (4.5), we obtain a boundary integral equation for the unknown density ψ:

$$\theta Y[\mathit{\Gamma},\psi](x_0) + (1-\theta)\overline{W}[\mathit{\Gamma},\psi](x_0) + (1-\theta)\alpha\psi(x_0) = v(x_0) \qquad \forall x_0 \in \mathit{\Gamma}. \tag{6.8}$$

From this we can deduce that (u,p) defined in (6.1) will be the unique solution of (2.1) if and only if ψ satisfies (6.8). This equation can be written in the standard form

$$\int_{\mathit{\Gamma}} K_\theta(x,y)\psi(y)\,dA_y + c_\theta\psi(x) = v(x) \qquad \forall x \in \mathit{\Gamma}, \tag{6.9}$$

where x_0 has been replaced by x for convenience, $c_\theta = (1-\theta)\alpha$, and

$$\begin{aligned} K^{ij}_\theta(x,y) \quad = \quad & \theta U^{\mathrm{PF}}_{ij}(x,x_*) + \theta U^{\mathrm{ROT}}_{il}(x,x_*)\varepsilon_{lpj}(y_p - x_{*p}) \\ & \qquad\qquad + (1-\theta)U^{\mathrm{STR}}_{ijl}(x,y)\nu_l(y). \end{aligned} \tag{6.10}$$

Remarks 6.1.

1. As before, assuming Γ is a Lyapunov surface, the kernel function $K_\theta(x,y)$ can be shown to be weakly singular. Thus the solvability of the linear integral equation (6.9) can be assessed via the Fredholm theory [19, 23]. Notice that $c_\theta = 0$ when $\theta = 1$, and $c_\theta \neq 0$ when $\theta \in [0,1)$. Thus (6.9) is a Fredholm equation of the first kind when $\theta = 1$ and of the second kind when $\theta \in [0,1)$.
2. The main idea in Power and Miranda [28] can be described intuitively as follows. A double-layer potential is deficient in the sense that it can only produce flows with zero resultant force and torque on Γ. Thus, in view of the flow properties outlined in (3.7) and (3.13), the enhancement of a double-layer potential with point-force and rotlet solutions should produce flows with arbitrary resultant force and torque on Γ.
3. The intuitive arguments above are made rigorous by the results outlined below. They show that the representation in (6.1) with $\theta \in (0,1)$ is complete in the sense that it can represent arbitrary flows and stable in the sense that the density ψ depends continuously on the boundary data v.

6.2. Solvability result. The following is a slight generalization of the solvability result given in Power and Miranda [28].

THEOREM 6.1 (see [28]). *Assume Γ is a closed, bounded Lyapunov surface, and let $x_* \in B$ be arbitrary. If $\theta \in (0,1)$, then* (6.9) *possesses a unique continuous solution ψ for any continuous boundary data v.*

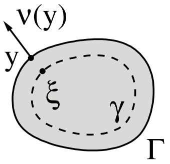
Thus arbitrary solutions of the exterior Stokes boundary-value problem (2.1) can be represented in the form (6.1) with a unique density ψ for each $x_* \in B$ and $\theta \in (0,1)$. The presence of the double-layer potential in (6.1) ensures that the representation is stable. In particular, because (6.9) is a uniquely solvable Fredholm equation of the second kind, the linear operator in (6.9) has a finite condition number and the density ψ depends continuously on the data v. The presence of the point-force and rotlet functions in (6.1) ensures that the representation is complete. In particular, these two singular solutions complete the deficient range associated with the double-layer potential. The smoothness properties of the density ψ depend on those of the surface Γ and the data v.

Remarks 6.2.

1. Aside from the restriction of solvability, the parameters θ and x_* are arbitrary and can be exploited. For example, $\theta \in (0,1)$ and $x_* \in B$ might be chosen by some means to optimize the conditioning of the linear operator in (6.9).
2. The boundary integral equation in (6.9) can be described as being singularity-free. The singularity in the point-force and rotlet contributions is avoided because their poles are contained in the body domain B. Moreover, the singularity in the kernel of the double-layer potential can be removed in a simple, standard way by employing a well-known integral identity [11, 28, 30, 31] (see section 8).
3. The above results hold for bodies of arbitrary shape. For certain types of bodies, for example convex or star-shaped bodies, there are various reasonable choices for the point $x_* \in B$. The center of volume is one obvious choice. In contrast, for other types of bodies, for example, toroidal or knotted tubular bodies, there seems to be no natural choice for the point $x_* \in B$. Motivated by this latter class of bodies, we investigate an alternative formulation.

7. A new formulation. Here we introduce a new boundary integral formulation of (2.1) which combines the strengths of the Power and Miranda and the Hebeker formulations. The new formulation is stable, complete, singularity-free, and natural for bodies of complicated shape and topology. In what follows $\mathit{\Gamma}$ is an arbitrary closed, bounded surface with interior domain B and exterior domain B_e, as described in section 2.

7.1. Formulation. Let γ be a surface parallel to $\mathit{\Gamma}$ offset toward B by a distance $\phi \geq 0$. In particular, γ is the image of the map $\xi = \zeta(y) : \mathit{\Gamma} \to \mathbb{R}^3$ defined by



$$\xi = y - \phi\, \nu(y). \tag{7.1}$$

By virtue of the fact that $\mathit{\Gamma}$ is a Lyapunov surface, it follows that the map $\zeta : \mathit{\Gamma} \to \gamma$ is continuous and one-to-one for all $\phi \in [0, \phi_{\mathit{\Gamma}})$, where $\phi_{\mathit{\Gamma}}$ is a positive constant. In the absence of any global obstructions, we have $\phi_{\mathit{\Gamma}} = 1/\kappa_{\mathit{\Gamma}}$, where $\kappa_{\mathit{\Gamma}}$ is the maximum of the signed principal curvatures of $\mathit{\Gamma}$ [27]. Here we use the convention that the curvature is positive when $\mathit{\Gamma}$ curves away from its outward unit normal ν. As a consequence, the principal curvatures are the eigenvalues of the gradient of ν (not $-\nu$) restricted to the tangent plane. From the geometry of parallel surfaces we have the following relations for all $y \in \mathit{\Gamma}$, $\xi = \zeta(y) \in \gamma$, and $\phi \in [0, \phi_{\mathit{\Gamma}})$ [27]:

$$n(\xi) = \nu(y), \quad dA_\xi = J^\phi(y)\, dA_y, \quad J^\phi(y) = 1 - 2\phi\kappa^m(y) + \phi^2\kappa^g(y). \tag{7.2}$$

Here n is the outward unit normal on γ, dA_ξ and dA_y are area elements on γ and $\mathit{\Gamma}$, and κ^m and κ^g are the mean and Gaussian curvatures of $\mathit{\Gamma}$. For $\phi \in [0, \phi_{\mathit{\Gamma}})$ we denote the inverse of $\xi = \zeta(y)$ by $y = \varphi(\xi)$. In view of (7.1) and $(7.2)_1$ we have $y = \xi + \phi n(\xi)$.

Given an arbitrary density $\psi : \mathit{\Gamma} \to \mathbb{R}^3$ and number $\theta \in [0, 1]$, define $u : B_e \to \mathbb{R}^3$ and $p : B_e \to \mathbb{R}$ by

$$u = \theta V[\gamma, \psi \circ \varphi] + (1 - \theta) W[\mathit{\Gamma}, \psi], \quad p = \theta P_V[\gamma, \psi \circ \varphi] + (1 - \theta) P_W[\mathit{\Gamma}, \psi]. \tag{7.3}$$

Notice that the double-layer potentials are defined on the surface $\mathit{\Gamma}$ with density ψ, while the single-layer potentials are defined on the parallel surface γ with density $\psi \circ \varphi$. In particular, the two types of potentials are defined on different surfaces but involve only one arbitrary density.

By properties of the single- and double-layer potentials, the fields (u, p) are smooth at each $x \in B_e$ and satisfy the Stokes equations $(2.1)_{1,2,4}$ at each $x \in B_e$. The stress field $\sigma : B_e \to \mathbb{R}^{3\times 3}$ associated with (u, p) is given by

$$\sigma = \theta \mathit{\Sigma}_V[\gamma, \psi \circ \varphi] + (1 - \theta) \mathit{\Sigma}_W[\mathit{\Gamma}, \psi], \tag{7.4}$$

and the resultant force F, torque T about an arbitrary point c, and volume flow rate Q associated with $\mathit{\Gamma}$ are

$$F = \theta F_V[\gamma, \psi \circ \varphi], \quad T = \theta T_V[\gamma, \psi \circ \varphi], \quad Q = (1 - \theta) Q_W[\mathit{\Gamma}, \psi]. \tag{7.5}$$

Here we have used linearity and the flow properties of the single- and double-layer potentials outlined in section 4.4.

In order for (u, p) to provide the unique solution of the exterior Stokes boundary-value problem (2.1), the boundary condition $(2.1)_3$ must be satisfied. In particular, given $v : \mathit{\Gamma} \to \mathbb{R}^3$, we require

$$\lim_{\substack{x \to x_0 \\ x \in B_e}} u(x) = v(x_0) \quad \forall x_0 \in \mathit{\Gamma}. \tag{7.6}$$

Substituting for u from (7.3) and using the limit relation in (4.5), we obtain a boundary integral equation for the unknown density ψ:

$$\theta V[\gamma, \psi \circ \varphi](x_0) + (1-\theta)\overline{W}[\mathit{\Gamma}, \psi](x_0) + (1-\theta)\alpha\psi(x_0) = v(x_0) \quad \forall x_0 \in \mathit{\Gamma}. \tag{7.7}$$

From this we can deduce that (u, p) defined in (7.3) will be the unique solution of (2.1) if and only if ψ satisfies (7.7). By definition of the single- and double-layer potentials, this equation can be written in integral form as

$$\begin{aligned} &\theta \int_\gamma U_{ij}^{\mathrm{PF}}(x,\xi)\psi_j(\varphi(\xi))\, dA_\xi \\ &\qquad + (1-\theta)\int_{\mathit{\Gamma}} U_{ijl}^{\mathrm{STR}}(x,y)\psi_j(y)\nu_l(y)\, dA_y + c_\theta\psi_i(x) = v_i(x) \quad \forall x \in \mathit{\Gamma}, \end{aligned} \tag{7.8}$$

where x_0 has been replaced by x for convenience and $c_\theta = (1-\theta)\alpha$. By performing a change of variable in the first integral, this equation can then be put into the standard form

$$\int_{\mathit{\Gamma}} K_\theta(x,y)\psi(y)\, dA_y + c_\theta\psi(x) = v(x) \qquad \forall x \in \mathit{\Gamma}, \tag{7.9}$$

where

$$K_\theta^{ij}(x,y) = \theta J^\phi(y) U_{ij}^{\mathrm{PF}}(x,\zeta(y)) + (1-\theta)U_{ijl}^{\mathrm{STR}}(x,y)\nu_l(y). \tag{7.10}$$

Remarks 7.1.

1. In all three formulations the kernel function $K_\theta(x,y)$ can be described as the positive linear combination of a double-layer kernel and a range completion term. In the Hebeker formulation (5.7), the completion term is an unbounded single-layer kernel. In the Power and Miranda formulation (6.10), the completion term is the sum of a point-force and a rotlet kernel, both of which are bounded and dependent on a point $x_* \in B$. In the new formulation (7.10), the completion term can be interpreted as a regularized single-layer kernel, where $\phi \geq 0$ is the regularization parameter. The regularized single-layer kernel is bounded when $\phi > 0$ and unbounded exactly as in the Hebeker formulation when $\phi = 0$.
2. Assuming $\mathit{\Gamma}$ is a Lyapunov surface, the kernel function $K_\theta(x,y)$ in (7.10) can be shown to be weakly singular. Thus the solvability of the linear integral equation (7.9) can be assessed via the Fredholm theory [19, 23]. Notice that $c_\theta = 0$ when $\theta = 1$ and $c_\theta \neq 0$ when $\theta \in [0,1)$. Thus (7.9) is a Fredholm equation of the first kind when $\theta = 1$ and of the second kind when $\theta \in [0,1)$.
3. As outlined below, the representation in (7.3) with $\theta \in (0,1)$ is complete in the sense that it can represent arbitrary flows and stable in the sense that the density ψ depends continuously on the boundary data v.

7.2. Solvability result. The following result establishes the solvability of the integral equation (7.9), or, equivalently, (7.7). Its proof is given in section 7.3 below.

THEOREM 7.1. *Assume $\mathit{\Gamma}$ is a closed, bounded Lyapunov surface, and let γ be a surface parallel to $\mathit{\Gamma}$ offset toward B by a distance $\phi \in [0, \phi_{\mathit{\Gamma}})$. If $\theta \in (0,1)$, then* (7.9) *possesses a unique continuous solution ψ for any continuous boundary data v.*

Thus arbitrary solutions of the exterior Stokes boundary-value problem (2.1) can be represented in the form (7.3) with a unique density ψ for each $\theta \in (0,1)$ and

$\phi \in [0, \phi_\Gamma)$. The presence of the double-layer potential in (7.3) ensures that the representation is stable. In particular, because (7.9) is a uniquely solvable Fredholm equation of the second kind, the linear operator in (7.9) has a finite condition number, and the density ψ depends continuously on the data v. The presence of the single-layer potential in (7.3) ensures that the representation is complete. In particular, the single-layer potential on the parallel surface γ completes the deficient range associated with the double-layer potential on the surface Γ. The smoothness properties of the density ψ depend on those of the surface Γ and the data v.

Remarks 7.2.

1. Aside from the restriction of solvability, the parameters θ and ϕ are arbitrary and can be exploited. For example, $\theta \in (0, 1)$ and $\phi \in [0, \phi_\Gamma)$ might be chosen by some means to optimize the conditioning of the linear operator in (7.9).
2. Just like the Power and Miranda formulation, the current boundary integral equation in (7.9) is singularity-free in the case when $\phi > 0$. The singularity in the single-layer potential is removed by virtue of the parallel surface. The singularity in the kernel of the double-layer potential can be removed in a simple, standard way by employing a well-known integral identity [11, 28, 30, 31] (see section 8).
3. Just like the Hebeker formulation, the current boundary integral formulation is natural for bodies of arbitrary shape. It seems particularly well suited for long, uniform, tubular bodies with complicated topology. In this case, the maximum offset distance ϕ_Γ can be explicitly identified as the tube radius, and Γ and γ would be parallel tubular surfaces of different radii centered on the same axial curve. In general, however, an explicit characterization of ϕ_Γ is not necessary, and the formulation is valid for any type of body.
4. All three formulations can be viewed as extensions to Stokes flow of ideas developed in classic potential theory. The idea of taking a linear combination of single- and double-layer potentials was considered in the work of Günter [13], and the idea of moving the single-layer potential to an inner surface, or limit thereof, was suggested in the work of Mikhlin [23]. (Mikhlin explicitly considered an inner point-source, which can be interpreted as the limit of a single-layer potential as the inner surface is squeezed to a point.) Other generalized formulations could also be considered. For example, the Power and Miranda formulation can be generalized by using a continuous distribution of stokeslet and rotlet singularities over an inner surface.

7.3. Proof of Theorem 7.1. Assume $\theta \in (0, 1)$ and consider the homogeneous version of (7.7). Replacing ψ by ψ^h for notational convenience, we have

$$\theta V[\gamma, \psi^h \circ \varphi](x_0) + (1-\theta)\overline{W}[\Gamma, \psi^h](x_0) + (1-\theta)\alpha\psi^h(x_0) = 0 \quad \forall x_0 \in \Gamma. \tag{7.11}$$

According to the Fredholm theory [19, 23], if (7.11) possesses only the trivial solution $\psi^h = 0$, then (7.7) possesses a unique continuous solution ψ for any continuous data v. To show that $\psi^h = 0$ is the only solution of (7.11), we proceed in four steps.

(1) Let ψ^h be an arbitrary solution of (7.11), and introduce fields $(u^{(1)}, p^{(1)})$ and $(u^{(2)}, p^{(2)})$ by

$$\begin{aligned} u^{(1)} &= (1-\theta)W[\Gamma, \psi^h], \quad & p^{(1)} &= (1-\theta)P_W[\Gamma, \psi^h], \\ u^{(2)} &= -\theta V[\gamma, \psi^h \circ \varphi], \quad & p^{(2)} &= -\theta P_V[\gamma, \psi^h \circ \varphi]. \end{aligned} \tag{7.12}$$

Then $(u^{(1)}, p^{(1)})$ and $(u^{(2)}, p^{(2)})$ satisfy the homogeneous Stokes equations $(2.1)_{1,2,4}$

in B_e. Moreover, from (7.11) and the limit relation for $W[\mathit{\Gamma}, \psi^h]$ in (4.5), we have

$$\lim_{\substack{x \to x_0 \\ x \in B_e}} u^{(1)}(x) - u^{(2)}(x) = 0 \quad \forall x_0 \in \mathit{\Gamma}. \tag{7.13}$$

Thus $u^{(1)} = u^{(2)}$ on $\mathit{\Gamma}$, and by uniqueness of solutions of the boundary-value problem (2.1), we have $(u^{(1)}, p^{(1)}) = (u^{(2)}, p^{(2)})$ in B_e. Furthermore, by properties of the single- and double-layer potentials defined in (4.1) and (4.2), we have $u^{(1)} = O(|x|^{-2})$ and $u^{(2)} = O(|x|^{-1})$ as $|x| \to \infty$, and $p^{(1)} = O(|x|^{-3})$ and $p^{(2)} = O(|x|^{-2})$. Thus we deduce

$$u^{(1)} = u^{(2)} = 0 \quad \text{and} \quad p^{(1)} = p^{(2)} = 0 \quad \forall x \in B_e. \tag{7.14}$$

(2) Since $u^{(1)} = 0$ in B_e and $1 - \theta \neq 0$, we deduce from (7.12) that $W[\mathit{\Gamma}, \psi^h] = 0$ in B_e, which implies

$$\lim_{\substack{x \to x_0 \\ x \in B_e}} W[\mathit{\Gamma}, \psi^h](x) = 0 \qquad \forall x_0 \in \mathit{\Gamma}. \tag{7.15}$$

Using the limit relation in (4.5), we get

$$\overline{W}[\mathit{\Gamma}, \psi^h](x_0) + \alpha \psi^h(x_0) = 0 \quad \forall x_0 \in \mathit{\Gamma}. \tag{7.16}$$

By well-known properties of the double-layer potential [20, 26], the above equation possesses exactly six independent eigenfunctions $\psi^{h,(1)}, \ldots, \psi^{h,(6)}$ defined for $x \in \mathit{\Gamma}$ by

$$\begin{aligned} \psi_i^{h,(a)}(x) &= \delta_{ia}, & a &= 1, 2, 3, \\ \psi_i^{h,(a)}(x) &= \varepsilon_{ij(a-3)} x_j, & a &= 4, 5, 6. \end{aligned} \tag{7.17}$$

Thus every solution ψ^h of (7.11) satisfies (7.16) and must necessarily be of the form

$$\psi^h(x) = \sum_{a=1}^{6} c_a \psi^{h,(a)}(x), \tag{7.18}$$

where $c_1, \ldots, c_6$ are arbitrary constants.

(3) Since $u^{(2)} = 0$ and $p^{(2)} = 0$ in B_e and $\theta \neq 0$, we deduce from (7.12) that $V[\gamma, \psi^h \circ \varphi] = 0$ and $P_V[\gamma, \psi^h \circ \varphi] = 0$ in B_e. Thus the resultant force and torque, about an arbitrary point q, exerted on $\mathit{\Gamma}$ by the exterior single-layer flow $(V[\gamma, \psi^h \circ \varphi], P_V[\gamma, \psi^h \circ \varphi])$ must vanish. By properties of the single-layer potentials outlined in section 4.4, and considering that $\gamma \subset \mathit{\Gamma}_{\text{int}}$ when $\phi > 0$ and $\gamma = \mathit{\Gamma}$ when $\phi = 0$, we find in both cases that

$$\begin{aligned} F_V[\gamma, \psi^h \circ \varphi] &= -8\pi \int_\gamma \psi^h(\varphi(\xi))\, dA_\xi = 0, \\ T_V[\gamma, \psi^h \circ \varphi] &= -8\pi \int_\gamma (\xi - q) \times \psi^h(\varphi(\xi))\, dA_\xi = 0. \end{aligned} \tag{7.19}$$

Dividing by -8π and substituting for ψ^h using (7.18) and (7.17), we find that the above equations yield a linear system for $c = (c_1, \ldots, c_6)$ of the form

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} c = 0, \tag{7.20}$$

where $A, B, C, D \in \mathbb{R}^{3\times 3}$ are defined by

$$\begin{aligned} A_{ij} &= \int_\gamma \delta_{ij}\, dA_\xi, \quad B_{ik} = \int_\gamma \varepsilon_{ijk}\varphi_j(\xi)\, dA_\xi, \\ C_{ij} &= \int_\gamma \varepsilon_{ipj}(\xi_p - q_p)\, dA_\xi, \quad D_{ik} = \int_\gamma \varepsilon_{ipl}\varepsilon_{ljk}(\xi_p - q_p)\varphi_j(\xi)\, dA_\xi. \end{aligned} \tag{7.21}$$

(4) For convenience, let the torque reference point q be the centroid of γ, and assume without loss of generality that $q = 0$. Then $C_{ij} = 0$ and (7.20) will possess only the trivial solution provided that the matrix D_{ik} is invertible. Notice that the matrix A_{ij} is always invertible since γ has positive measure. With $q = 0$ we have

$$D_{ik} = \int_\gamma \varepsilon_{ipl}\varepsilon_{ljk}\xi_p\varphi_j(\xi)\, dA_\xi. \tag{7.22}$$

Substituting $\varphi(\xi) = \xi + \phi n(\xi)$, where n is the outward unit normal field on γ (see section 7.1), and using the standard permutation symbol identity $\varepsilon_{ipl}\varepsilon_{ljk} = \delta_{ij}\delta_{pk} - \delta_{ik}\delta_{pj}$, we obtain

$$D_{ik} = \int_\gamma \xi_i\xi_k - \delta_{ik}\xi_j\xi_j\, dA_\xi + \phi\int_\gamma \xi_k n_i\, dA_\xi - \phi\delta_{ik}\int_\gamma \xi_j n_j\, dA_\xi. \tag{7.23}$$

Applying the divergence theorem to the last two integrals in the above equation, we get, after straightforward simplification,

$$D_{ik} = -[G_{ik} + 2\phi\,\mathrm{vol}(\gamma_{\mathrm{int}})\delta_{ik}]. \tag{7.24}$$

Here $\mathrm{vol}(\gamma_{\mathrm{int}}) > 0$ is the volume of the interior domain enclosed by γ and $G_{ik} = \int_\gamma \delta_{ik}|\xi|^2 - \xi_i\xi_k\, dA_\xi$ is the symmetric, positive-definite, second moment tensor associated with γ. Since D_{ik} is invertible for any $\phi \geq 0$, we deduce that (7.20) admits only the trivial solution $c = 0$. Combining this with (7.18), we deduce that (7.11) admits only the trivial solution $\psi^h = 0$. This completes the proof of Theorem 7.1.

8. Nyström approximation.

In this section we describe a numerical method for the formulation presented in section 7. We outline a singularity-free formulation of the boundary integral equation for the unknown density, discretize it using a straightforward Nyström method with an arbitrary quadrature rule, and introduce corresponding discretizations for various flow quantities of interest.

8.1. Singularity-free formulation.

Given arbitrary parameters $\theta \in (0, 1)$ and $\phi \in (0, \phi_\Gamma)$, the integral equation (7.9) can be written in the convenient form

$$\begin{aligned} \theta\int_\Gamma G(x,y)\psi(y)\, dA_y + (1-\theta)\int_\Gamma H(x,y)\psi(y)\, dA_y \\ + (1-\theta)\alpha\psi(x) = v(x) \quad \forall x \in \Gamma, \end{aligned} \tag{8.1}$$

where $G_{ij}(x,y) = J^\phi(y)U_{ij}^{\mathrm{PF}}(x,\zeta(y))$ is the regularized, bounded, single-layer kernel and $H_{ij}(x,y) = U_{ijl}^{\mathrm{STR}}(x,y)\nu_l(y)$ is the standard, weakly-singular, double-layer kernel. The singularity in $H(x,y)$ can be avoided in a simple way by exploiting the double-layer identity [11, 28, 30, 31]

$$\int_\Gamma H_{ij}(x,y)\, dA_y = -\alpha\delta_{ij} \quad \forall x \in \Gamma. \tag{8.2}$$

In particular, substitution of (8.2) into (8.1) gives

$$\begin{aligned}(8.3)\qquad &\theta \int_{\Gamma} G(x,y)\psi(y)\,dA_y \\ &\qquad + (1-\theta)\int_{\Gamma} H(x,y)[\psi(y)-\psi(x)]\,dA_y = v(x) \quad \forall x \in \Gamma.\end{aligned}$$

Assuming ν and ψ are Lipschitz continuous and Γ is a Lyapunov surface, it can be shown that the functions $G(x,y)\psi(y)$ and $H(x,y)[\psi(y)-\psi(x)]$ are uniformly bounded for all x and y on Γ. Thus (8.3) is singularity-free and can be discretized by Nyström methods.

Remarks 8.1.

1. Following standard practice [28, 30, 31], we define $H(x,y)[\psi(y)-\psi(x)]$ to be zero when $y = x$. This modification does not alter the value of the integral and is necessary to obtain a well-defined Nyström discretization, which requires a value for this function for arbitrary x and y. The results in [31] show that Nyström methods defined using this practice are, in general, convergent. However, there is generally an upper bound on the order of convergence of these methods because of the above modification.
2. There is some freedom in the treatment of the first term in (8.3). When written as an integral over Γ as above, the kernel function $G(x,y)$ contains the factor $J^{\phi}(y)$, which depends on the curvature of Γ (see section 7.1). By a change of variable, this term could also be written as an integral over the parallel surface γ. In this case, the curvature factor disappears, but an explicit parameterization of γ becomes necessary.

8.2. Approximation of integral equation. We suppose Γ can be decomposed into a union of nonoverlapping patches Γ_p, $p = 1,\dots,M_p$, where each patch is the image of a smooth map $y = \chi_p(s,t) : D_p \to \mathbb{R}^3$, and each D_p is a domain in $\mathbb{R}^2$. By subdividing each domain D_p into nonoverlapping subdomains D_p^e, $e = 1,\dots,M_e$, we decompose each patch Γ_p into curved, nonoverlapping patch elements Γ_p^e. In each patch element we introduce quadrature points $y_{p,e,q}$ and weights $W_{p,e,q}$, $q = 1,\dots,M_q$, such that

$$(8.4)\qquad \int_{\Gamma_p^e} f(y)\,dA_y = \int_{D_p^e} f(\chi_p(s,t))J_p(s,t)\,ds\,dt \approx \sum_{q=1}^{M_q} f(y_{p,e,q})W_{p,e,q}.$$

Here J_p is the Jacobian associated with the patch parameterization χ_p, which is assumed to be included in the weights $W_{p,e,q}$.

Let $\psi_{p,e,q}$ be an approximation to $\psi(y_{p,e,q})$, and for convenience let a and b denote values of the multi-index (p,e,q). Then a Nyström discretization of (8.3) is

$$(8.5)\qquad \theta\sum_{b} G_{ab}\psi_b W_b + (1-\theta)\sum_{b\neq a} H_{ab}[\psi_b-\psi_a]W_b = v_a \quad \forall a,$$

where $G_{ab} = G(x_a,y_b)$, $H_{ab} = H(x_a,y_b)$, and $v_a = v(x_a)$. Here the product $H_{ab}[\psi_b - \psi_a]$ has been set equal to zero when $b = a$. The above equation can be written in the standard form

$$(8.6)\qquad \sum_{b} A_{ab}\psi_b = v_a \quad \forall a,$$

where $A_{ab} \in \mathbb{R}^{3\times 3}$ are defined by

$$A_{ab} = \begin{cases} \theta G_{ab} W_b + (1-\theta) H_{ab} W_b, & a \neq b, \\ \theta G_{aa} W_a - (1-\theta) \sum_{c \neq a} H_{ac} W_c, & a = b. \end{cases} \tag{8.7}$$

Equation (8.6) is a linear system of algebraic equations for the approximate density values ψ_b at the quadrature points y_b. This system is dense and nonsymmetric and can be solved using any suitable numerical technique.

8.3. Approximation of flow quantities. Various flow quantities of interest take the form of an integral of ψ over Γ. For example, from (7.5), the resultant force and torque on Γ about an arbitrary point c are given by

$$F = -8\pi\theta \int_\gamma \psi(\varphi(\xi))\, dA_\xi, \quad T = -8\pi\theta \int_\gamma (\xi - c) \times \psi(\varphi(\xi))\, dA_\xi. \tag{8.8}$$

After a change of variable (see section 7.1), these integrals can be transformed from the parallel surface γ to the body surface Γ to obtain

$$F = -8\pi\theta \int_\Gamma J^\phi(y) \psi(y)\, dA_y, \quad T = -8\pi\theta \int_\Gamma J^\phi(y)(\zeta(y) - c) \times \psi(y)\, dA_y. \tag{8.9}$$

By discretizing these integrals using the same quadrature points and weights as before, we get the approximations

$$F^{\text{approx}} = -8\pi\theta \sum_b J_b^\phi \psi_b W_b, \quad T^{\text{approx}} = -8\pi\theta \sum_b J_b^\phi (\zeta_b - c) \times \psi_b W_b. \tag{8.10}$$

An approximation to the volume flow rate Q associated with Γ can be obtained in a similar manner.

9. Numerical experiments. Here we present results from numerical experiments on three different bodies: a sphere, torus, and helical tube with hemispherical endcaps. For one or more prescribed motions of each body, we computed the resultant force and torque about the origin of a body-fixed frame and examined various measures of convergence.

9.1. Methods. Following the general procedure outlined above, we decomposed the surface of each body into nonoverlapping patches Γ_p, each parameterized over a rectangular domain D_p. We subdivided each patch into curved, quadrilateral elements Γ_p^e, and in each element we used an $m \times m$ tensor product Gauss–Legendre quadrature rule, with order of accuracy $r_{\text{abs}} = 2m$ on absolute errors. For the sphere we employed six patches based on stereographic projection from the faces of a bounding cube. For the torus we employed a single patch based on an explicit parameterization of the axial curve. For the helical tube we employed multiple patches based on explicit parameterizations of the axial curve and endcaps.

The resultant force F and torque T on each body were computed by solving the linear algebraic system (8.6) of size $(3M_pM_eM_q) \times (3M_pM_eM_q)$. Because this system is nonsymmetric and was observed to be well conditioned, we used the GMRES iterative solver implemented in MATLAB with no preconditioning and a residual tolerance of 10^{-12}. Using the solution of (8.6), we computed approximations to F and T according to (8.10). The total number of quadrature points, $M_pM_eM_q$, was varied up to a maximum value of 4000 to 8000 depending on the example. All computations were performed with the parameter values $\theta = 1/2$ and $\phi/\phi_\Gamma = 1/2$, where ϕ_Γ is the maximum offset distance for the parallel surface associated with Γ.

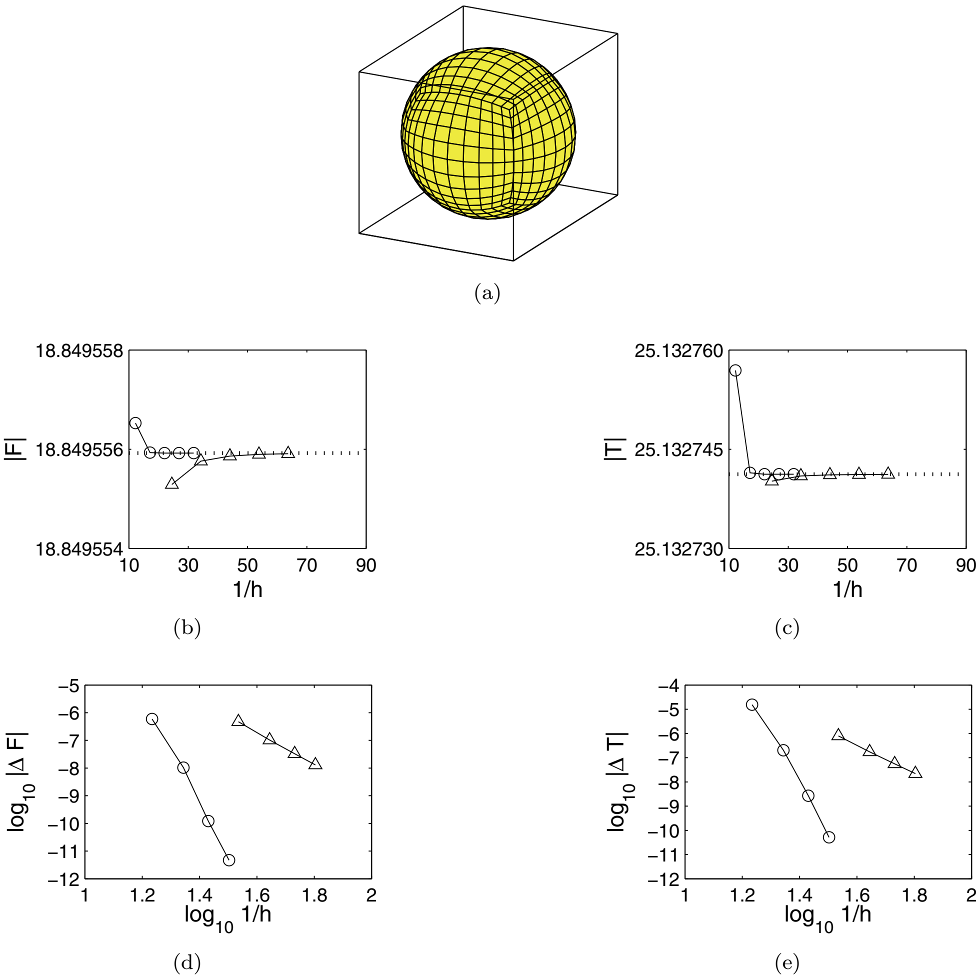


FIG. 9.1. *Convergence results for resultant force F and torque T on a sphere. Computations were performed with a sequence of meshes with element sizes h_k.* (a) *Sample mesh.* (b),(c) *Plots of $|F_{h_k}|$ and $|T_{h_k}|$ versus $1/h_k$ for the translational and rotational motion, respectively. The dotted horizontal lines indicate exact values.* (d),(e) *Plots of $\log_{10}|F_{h_k} - F_{h_{k-1}}|$ and $\log_{10}|T_{h_k} - T_{h_{k-1}}|$ versus $\log_{10}(1/h_k)$ for the translational and rotational motion, respectively. In all plots, triangles denote results for the 1×1 quadrature rule, and circles denote results for the 2×2 rule.*

9.2. Results. Figure 9.1 shows convergence results for the resultant force and torque about the origin on a sphere obtained with the 1×1 and 2×2 quadrature rules. The sphere had a radius $r = 1$ and was centered at the origin. For this surface, the maximum signed curvature is $\kappa_\Gamma = 1/r$, which gives a maximum offset distance of $\phi_\Gamma = r$. Results are given for two independent boundary conditions—translation along the x-axis with unit velocity, and rotation about the same axis with unit angular velocity. In these cases, exact values are known: $F = (-6\pi, 0, 0)$ and $T = (0, 0, 0)$ for the translational motion, and $F = (0, 0, 0)$ and $T = (-8\pi, 0, 0)$ for the rotational motion.

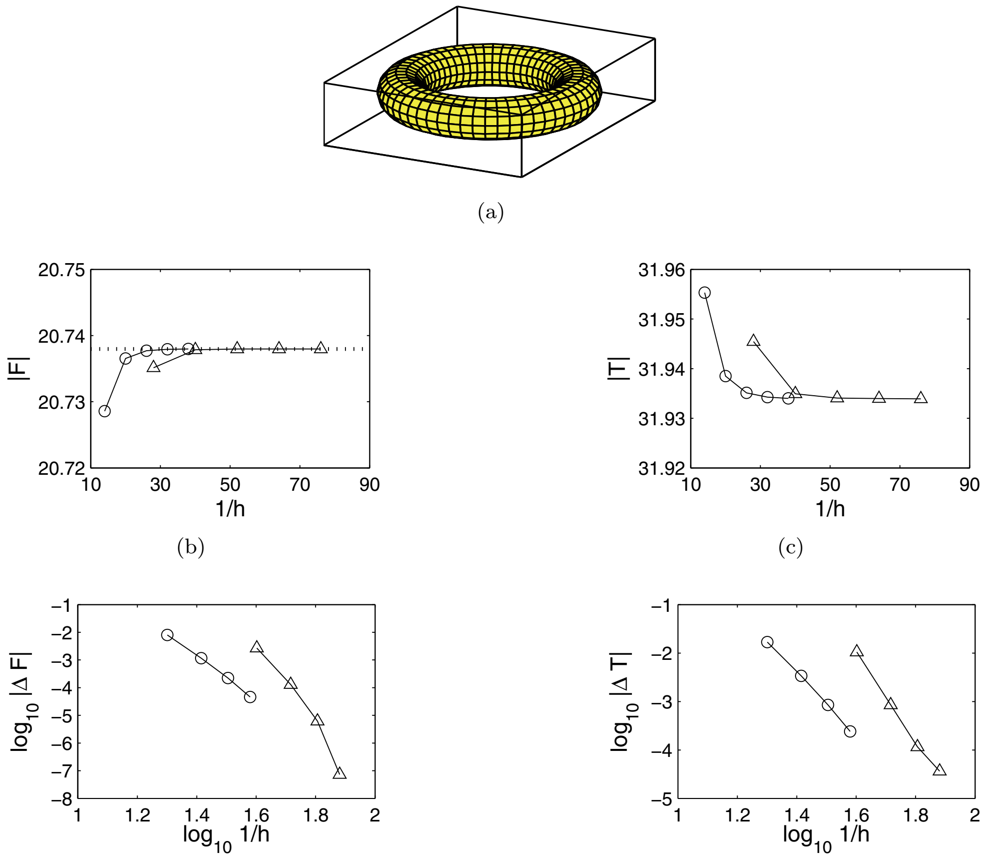
Plot (a) of Figure 9.1 illustrates the geometry and a sample mesh. In our computations, a sequence of five increasingly refined meshes were considered for each quadrature rule, with each mesh being relatively uniform. The meshes were chosen such that, at each stage in the sequence, the linear algebraic systems for the 1×1 and 2×2 quadrature rules were approximately the same size. The mesh shown in (a) is the coarsest used with the 1×1 rule. Plot (b) shows convergence results for the magnitude of F in the translational motion as a function of the element size parameter h, defined by $M_p M_e h^2 = 1$, where $M_p M_e$ is the total number of elements in a mesh. In particular, h is proportional to the average element size. Plot (c) shows similar convergence results for the magnitude of T in the rotational motion. In all computations, the appropriate entries in both F and T were found to be zero within machine precision for each type of motion. Thus the errors illustrated can be attributed to the appropriate nonzero components.

Plot (d) of Figure 9.1 shows the difference in the computed values of F between successive meshes as a function of h for the translational motion. Although an exact solution is available, we consider solution differences rather than absolute errors for purposes of later comparison. Plot (e) shows similar results for the difference in the computed values of T for the rotational motion. For an $m \times m$ Gauss–Legendre quadrature rule, the convergence rate r_{diff} for solution differences is expected to be $2m+1$, which is one order higher than the standard convergence rate r_{abs} for absolute errors. The plots show that the observed convergence rate for solution differences was significantly higher than expected. Considering both F and T, we have $5 \leq r_{\text{diff}} \leq 6$ for $m = 1$ and $19 \leq r_{\text{diff}} \leq 20$ for $m = 2$. On the finest two meshes used with the 2×2 rule, the relative change in F for the translational motion was of order 10^{-13}, and the relative change in T for the rotational motion was of order 10^{-12}.

Figure 9.2 shows convergence results for the resultant force and torque about the origin on a torus. The axial curve of the torus was a circle of radius $\rho = 1$ centered at the origin in the xy-plane, and the tube section was a circle of radius $r = \rho(1-\eta)/(1+\eta)$, where $\eta = \tanh^2(1)$. This value of the tube radius was chosen to compare results against an exact solution from [34]. For this surface, the maximum signed curvature is $\kappa_\Gamma = 1/r$, which gives a maximum offset distance of $\phi_\Gamma = r$. Results are given for two independent boundary conditions—translation along the z-axis with unit velocity, and rotation about the same axis with unit angular velocity. Symmetry implies that the force and torque have the form $F = (0, 0, F_z)$ and $T = (0, 0, 0)$ for the translational motion and $F = (0, 0, 0)$ and $T = (0, 0, T_z)$ for the rotational motion. For the translational motion, the force F_z has been characterized, and its approximate numerical value is $F_z = -20.7379$ [34]. For the rotational motion, the torque T_z has also been characterized [16], but its approximate numerical value does not appear to be well known.

Plots (a) through (e) of Figure 9.2 are analogous to the previous example. In our computations, we again found that the appropriate entries in both F and T were zero within machine precision for each type of motion. Moreover, the observed convergence rate for solution differences was again higher than expected. Considering both F and T, we have $9 \leq r_{\text{diff}} \leq 16$ for $m = 1$ and $6 \leq r_{\text{diff}} \leq 8$ for $m = 2$. Interestingly, for the range of meshes considered here, the 1×1 rule performed better than the 2×2 rule. On the finest two meshes used with the 1×1 rule, the relative change in F for the translational motion was of order 10^{-9}, and the relative change in T for the rotational motion was of order 10^{-6}.

Figure 9.3 shows convergence results for the resultant force and torque about the origin on a helical tube. The axial curve of the tube was a helical curve about the z-axis with radius $\rho = 2$, pitch $\lambda = 3$, and arclength $l = 2\pi$. The tube had uniform, circular cross-sections of radius $r = 0.2$ and hemispherical endcaps of the



(a)

(b)

(c)

(d)

(e)

FIG. 9.2. *Convergence results for resultant force F and torque T on a torus. Computations were performed with a sequence of meshes with element sizes h_k.* (a) *Sample mesh.* (b),(c) *Plots of* $|F_{h_k}|$ *and* $|T_{h_k}|$ *versus* $1/h_k$ *for the translational and rotational motion, respectively. The dotted horizontal line in* (b) *indicates an exact value.* (d),(e) *Plots of* $\log_{10}|F_{h_k} - F_{h_{k-1}}|$ *and* $\log_{10}|T_{h_k} - T_{h_{k-1}}|$ *versus* $\log_{10}(1/h_k)$ *for the translational and rotational motion, respectively. In all plots, triangles denote results for the* 1×1 *quadrature rule, and circles denote results for the* 2×2 *rule.*

same radius. These geometrical parameters were chosen so as to produce a tubular body of moderately high curvature. As with the torus, the maximum signed curvature is $\kappa_\Gamma = 1/r$, which gives a maximum offset distance of $\phi_\Gamma = r$. In contrast to the previous two examples, results are given for a single boundary condition—rotation about the x-axis with unit angular velocity. In this case, the resultant force and torque are not known exactly and are not known to have any special form.

Plots (a) through (e) of Figure 9.3 are analogous to the previous two examples, with the exception that only one type of motion is considered. For this single motion the force and torque were each found to possess three nonzero components, in contrast to the previous examples. The observed convergence rate for solution differences was again higher than expected. Considering both F and T, we have $3 \leq r_{\text{diff}} \leq 6$ for $m = 1$ and $8 \leq r_{\text{diff}} \leq 10$ for $m = 2$. For T we notice that the results from the 2×2 rule converge to a limiting value monotonically from below, whereas the results from the 1×1 rule converge nonmonotonically from above. On the finest two meshes used

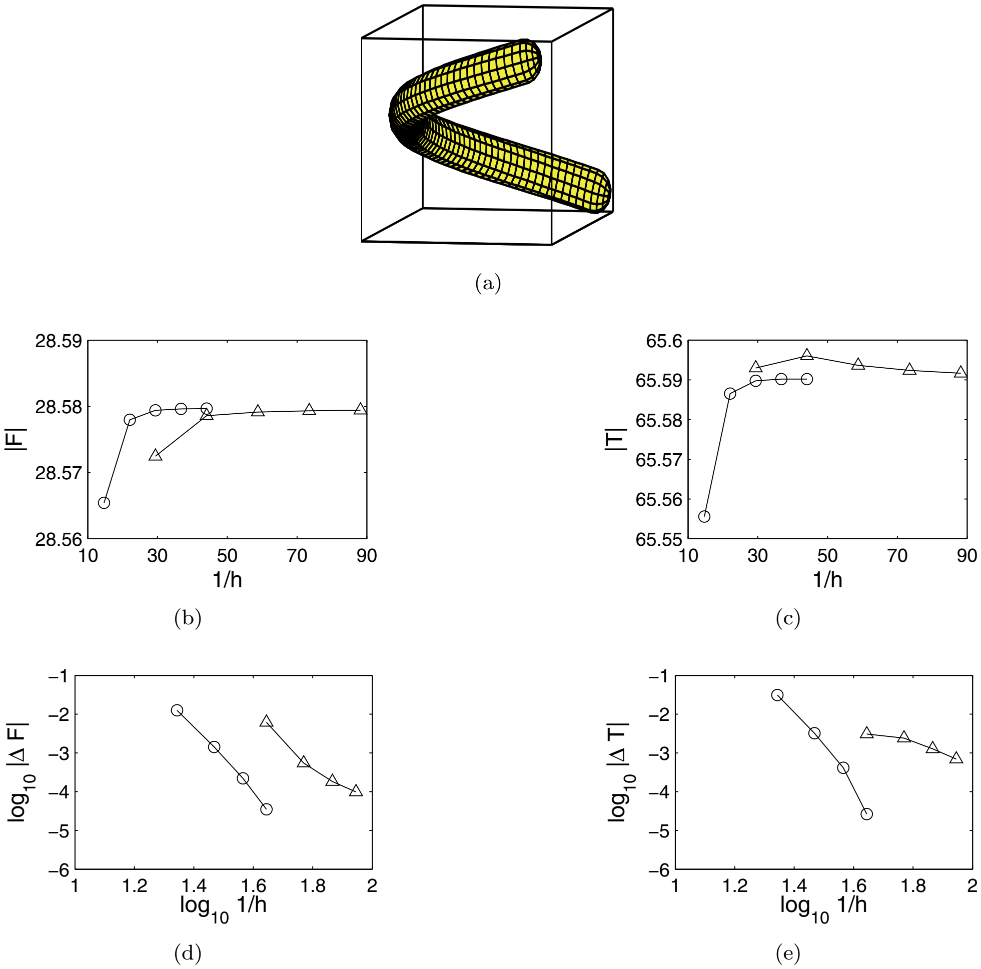


FIG. 9.3. *Convergence results for resultant force F and torque T on a helical tube. Computations were performed with a sequence of meshes with element sizes h_k.* (a) *Sample mesh.* (b),(c) *Plots of* $|F_{h_k}|$ *and* $|T_{h_k}|$ *versus* $1/h_k$. (d),(e) *Plots of* $\log_{10}|F_{h_k} - F_{h_{k-1}}|$ *and* $\log_{10}|T_{h_k} - T_{h_{k-1}}|$ *versus* $\log_{10}(1/h_k)$. *In all plots, triangles denote results for the* 1×1 *quadrature rule, and circles denote results for the* 2×2 *rule.*

with the 2×2 rule, the relative change in F was of order 10^{-6}, and the relative change in T was of order 10^{-7}.

9.3. Discussion. The examples outlined above suggest that the singularity-free boundary integral formulation introduced here leads to a viable numerical scheme for exterior Stokes flow problems. Issues associated with weakly singular integrals are avoided in a simple and efficient way without the need for product integration rules or specialized coordinate transformations and projections. In all three examples, the schemes exhibited convergence rates that were higher than expected and produced reasonably accurate results with reasonable meshes. For meshes of comparable size, the results for the torus and helical tube examples were less accurate than those for the sphere example. This is likely due to the relatively high curvature and more complicated shapes of the torus and helical tube. As can be expected, finer meshes are needed in these cases to achieve a level of accuracy similar to that for the sphere. The role of the parameters θ and ϕ in the conditioning and performance of these schemes for different classes of bodies will be investigated in a separate work.

Acknowledgments. The author thanks the reviewers for their helpful comments and the National Science Foundation for its generous support.

REFERENCES

[1] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin, *Wavelet-like bases for the fast solution of second-kind integral equations*, SIAM J. Sci. Comput., 14 (1993), pp. 159–184.

[2] K. E. Atkinson, *The numerical solution of Laplace's equation in three dimensions*, SIAM J. Numer. Anal., 19 (1982), pp. 263–274.

[3] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[4] G. K. Batchelor, *Slender-body theory for particles of arbitrary cross-section in Stokes flow*, J. Fluid Mech., 44 (1970), pp. 419–440.

[5] C. Brebbia, J. Telles, and L. Wrobel, *Boundary Element Techniques*, Springer-Verlag, New York, 1984.

[6] G. Chen and J. Zhou, *Boundary Element Methods*, Academic Press, New York, 1992.

[7] T. A. Dabros, *Singularity method for calculating hydrodynamic forces and particle velocities in low-Reynolds-number flows*, J. Fluid Mech., 156 (1985), pp. 1–21.

[8] M. G. Duffy, *Quadrature over a pyramid or cube of integrands with a singularity at a vertex*, SIAM J. Numer. Anal., 19 (1982), pp. 1260–1262.

[9] R. Finn, *On the exterior stationary problem for the Navier-Stokes equations and associated perturbation problems*, Arch. Ration. Mech. Anal., 19 (1965), pp. 363–406.

[10] M. Ganesh, I. G. Graham, and J. Sivaloganathan, *A new spectral boundary integral collocation method for three-dimensional potential problems*, SIAM J. Numer. Anal., 35 (1998), pp. 778–805.

[11] M. A. Goldberg and C. S. Chen, *Discrete Projection Methods for Integral Equations*, Computational Mechanics Publications, Billerica, MA, 1997.

[12] I. G. Graham and I. H. Sloan, *Fully discrete spectral boundary integral methods for Helmholtz problems on smooth closed surfaces in* $\mathbb{R}^3$, Numer. Math., 92 (2002), pp. 289–323.

[13] N. M. Günter, *Potential Theory and Its Applications to Basic Problems of Mathematical Physics*, Frederick Ungar Publishing, New York, 1967.

[14] F. K. Hebeker, *A boundary element method for Stokes equations in 3-D exterior domains*, in The Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., Academic Press, London, 1985, pp. 257–263.

[15] R. E. Johnson, *An improved slender-body theory for Stokes flow*, J. Fluid Mech., 99 (1980), pp. 411–431.

[16] R. P. Kanwal, *Slow steady rotation of axially symmetric bodies in a viscous fluid*, J. Fluid Mech., 10 (1961), pp. 17–24.

[17] J. B. Keller and S. I. Rubinow, *Slender-body theory for slow viscous flow*, J. Fluid Mech., 75 (1976), pp. 705–714.

[18] S. Kim and S. J. Karrila, *Microhydrodynamics*, Butterworth-Heinemann Publishing, Oxford, UK, 1991.

[19] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer-Verlag, New York, 1989.

[20] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, revised English ed., Gordon and Breach, New York, 1963.

[21] C. Lage and C. Schwab, *Wavelet Galerkin algorithms for boundary integral equations*, SIAM J. Sci. Comput., 20 (1999), pp. 2195–2222.

[22] P. A. Martin, *Multiple Scattering*, Encyclopedia Math. Appl. 107, Cambridge University Press, Cambridge, UK, 2006.

[23] S. G. Mikhlin, *Linear Integral Equations*, International Monographs on Advanced Mathematics and Physics, Hindustan Publishing Corporation, Delhi, 1960.

[24] S. G. Mikhlin, *Multidimensional Singular Integrals and Integral Equations*, International Series of Monographs in Pure and Applied Mathematics 83, Pergamon Press, Oxford, UK, 1965.

[25] S. G. Mikhlin and S. Prössdorf, *Singular Integral Operators*, Springer-Verlag, New York, 1986.

[26] F. K. G. Odqvist, *Über die randwertaufgaben der hydrodynamik zäher flüssigkeiten*, Math. Z., 32 (1930), pp. 329–375.

[27] B. O'Neill, *Elementary Differential Geometry*, Academic Press, New York, 1966.

[28] H. Power and G. Miranda, *Second kind integral equation formulation of Stokes' flows past a particle of arbitrary shape*, SIAM J. Appl. Math., 47 (1987), pp. 689–698.

[29] H. Power and L. C. Wrobel, *Boundary Integral Methods in Fluid Mechanics*, Computational Mechanics Publications, Billerica, MA, 1995.

[30] C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge University Press, Cambridge, UK, 1992.

[31] A. Rathsfeld, *Quadrature methods for 2D and 3D problems*, J. Comput. Appl. Math., 125 (2000), pp. 439–460.

[32] J. Tausch and J. White, *Multiscale bases for the sparse representation of boundary integral operators on complex geometry*, SIAM J. Sci. Comput., 24 (2003), pp. 1610–1629.

[33] L. Ying, G. Biros, and D. Zorin, *A high-order 3D boundary integral equation solver for elliptic PDEs in smooth domains*, J. Comput. Phys., 219 (2006), pp. 247–275.

[34] M. Zabarankin and P. Krokhmal, *Generalized analytic functions in 3D Stokes flows*, Quart. J. Mech. Appl. Math., 60 (2007), pp. 99–123.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 959–976

MULTIPARASITOID-HOST INTERACTIONS WITH EGG-LIMITED ENCOUNTER RATES[*]

RYUSUKE KON[†] AND SEBASTIAN J. SCHREIBER[‡]

Abstract. To address the contentious issue of multiple parasitoid introductions in classical biological control, a discrete-time model of multiparasitoid-host interactions that accounts for host density dependence and egg limitation is introduced and analyzed. For parasitoids that are egg limited but not search limited, the model is proven to exhibit four types of dynamics: host failure in which the host becomes extinct in the presence or absence of the parasitoids; parasitoid-driven extinction in which the parasitoid complex invariably drives the host extinct; host persistence; and conditional host persistence in which, depending on the initial ratios of host to parasitoid densities, the host is either driven extinct or persists. In the case of host persistence, the dynamics of the system are shown to be asymptotic to the dynamics of an appropriately defined one-dimensional difference equation. The results illustrate how the establishment of one or more parasitoids can facilitate the invasion of another parasitoid and how a complex of parasitoids can drive a host extinct despite every species in the complex being unable to do so. The effects of including search limitation are also explored.

Key words. host-parasitoid dynamics, permanence, extinction, Lyapunov exponents

AMS subject classifications. 92D25, 39A11, 37B25

DOI. 10.1137/080717006

1. Introduction.

Classical biological control is the introduction of natural enemies of a pest species with the goal of suppressing the abundance of the pest to a level at which it no longer causes economic damage [23]. For insect pests, control is often achieved by parasitoids: organisms, typically wasps and flies, whose young develop on and eventually kill their hosts. One of the earliest successes of biological control was with the cottony cushion scale, a pest that was devastating the developing California citrus industry in the late 1800s [4]. A predatory insect, the vedalia beetle (which functions as a parasitoid), and a parasitoid fly were introduced from Australia to control the cottony cushion scale. Within several years, these natural enemies suppressed this pest to very low densities, where they remain to this day when not disrupted by the use of broad-spectrum insecticides [23]. Since this pioneering project, there have been more than 3,600 intentional introductions of parasitoids to control more than 500 insect pests around the world [8]. Of these introductions, only 30% have resulted in the natural enemy establishing successfully, and of these only 36% have lead to substantial control of the targeted pest [8]. Consequently, there have been extensive theoretical and empirical efforts to understand what factors contribute to the success or failure of biological control programs. One particular contentious issue in these studies concerns whether or not the release of a single species or several species of natural enemy will lead to a lower host density. On the one hand, scientists have argued that it is essential to screen all natural enemies and release only the most effective species [32, 34, 5]. Others have argued that testing for the best

[*] Received by the editors February 27, 2008; accepted for publication (in revised form) August 28, 2008; published electronically January 14, 2009.

http://www.siam.org/journals/siap/69-4/71700.html

[†] Department of Biology, Faculty of Science, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan (kon-r@bio-math10.biology.kyushu-u.ac.jp).

[‡] Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, One Shields Avenue, Davis, CA 95616 (sschreiber@ucdavis.edu).

parasitoid species takes too much time and money and, consequently, have advocated releasing all available natural enemies [33, 15]. Theoretical studies have shown that whether multiple species introductions are advisable depends on the details of the biology [21, 18, 2, 29]. For instance, May and Hassell [21] argued that, in general, multiple parasitoid introductions result in greater suppression of the host than single parasitoid introductions. This conclusion, however, relied on the assumption that the parasitoid species aggregate independently of one another and independently of host density. Indeed, Kakehashi, Suzuki, and Iwasa [18] showed that single parasitoid introductions are more effective when both parasitoid species aggregate to the same regions of space. These theoretical studies assume that the parasitoids are search limited and not egg limited. Moreover, their analysis is typically limited to numerical simulations and, occasionally, equilibrium stability analysis. In contrast to these earlier studies, we analyze the global dynamics for multiparasitoid-host interactions when the parasitoids are egg limited but not search limited.

All parasitoids experience egg limitation to some degree [6, 12, 17, 19]. For instance, synovigenic parasitoids, which continuously produce eggs over their lifetime, experience egg limitation whenever the number of hosts they encounter in a day exceeds their daily production of eggs. In a field study, Heimpel and Rosenheim [12] caught and dissected 270 synovigenic parasitoids of the species *Aphelinidae aonidiae.* They found 18% of the dissected individuals had an egg load of zero and, consequently, were extremely egg limited. Several theoretical studies have examined the combined effects of egg limitation and search limitation on host-parasitoid dynamics [6, 10, 26, 30, 27, 28]. If one takes a broad view that egg limitation is a form of predator saturation, then it can be said that Rogers [26] was the first to consider egg limitation by translating Holling's type II functional response to a host encounter rate. Analyzing Roger's model, May and Hassell [10] found that egg limitation tends to destabilize host-parasitoid interactions. It was not until two decades later that the interaction of this destabilizing factor with a stabilizing factor (heterogeneity in the distribution of parasitoid attacks) was considered. Studying models without host self-regulation, Getz and Mills [6] found that stability of the host-parasitoid equilibrium requires parasitic attacks to be sufficiently aggregated and the intrinsic fitness of the parasitoid to exceed the intrinsic fitness of the host. Including host regulation, Schreiber found that parasitoids with aggregated attacks and sufficiently weak search limitation can suppress their hosts to extremely low densities and even drive them to extinction [27, 28]. None of these studies, however, considered how egg limitation influences multiparasitoid-host interactions. Given that classical biological control programs often involve the release of multiple parasitoid species, and that most parasitoids experience some degree of egg-limitation, an important facet of host-parasitoid dynamics remains to be understood.

To address this gap in our knowledge about host-parasitoid dynamics, we introduce and analyze a model of multiparasitoid-host interactions that accounts for egg-limitation. This model is presented in section 2. Using a simple change of variables introduced in [27], we provide in sections 3 and 4 a rather detailed analysis of the global dynamics for purely egg-limited parasitoids. In section 5, we examine the combined effects of weak search limitation and egg limitation. In section 6, we discuss the implications of our results for classical biological control.

2. Model. The discrete-time model describes the dynamics of host-parasitoid interactions with synchronized generations. The host of density N is subject to parasitism by n parasitoids of densities $P_1, \dots, P_n$. The fraction $g_i(E_i)$ of hosts escaping parasitism for species i depends on the host encounter rate E_i of parasitoid species i, a function of host and parasitoid density that is described in further detail below. The fraction of hosts escaping intraspecific density-dependent mortality is $f(N)$. Intraspecific density-dependent mortality is assumed to precede mortality due to parasitism (see, e.g., [13, 22, 27]). One interpretation of this assumption is that the parasitoids are koinobionts. Hence, the host continues to develop after being parasitized and experiences density-dependent mortality (via the survival function $f(\cdot)$) independent of parasitism. Hosts escaping parasitism and density-dependent mortality produce on average λ progeny that survive to the next generation. Following the approach taken by May and Hassell [21], we assume that there is a competitive hierarchy amongst parasitoid larvae: within a parasitized host, larvae from species i always outcompete larvae from species j whenever $j > i$. This assumption is appropriate for two types of interactions that are frequently found in host-parasitoid systems [21]. First, it applies when parasitoid species 1 attacks first, species 2 attacks second, etc. In these cases, the older parasitoid larvae are usually able to eliminate the younger competitors by physical suppression [7]. This situation is common when the parasitoid species attack different developmental stages of the host, e.g., parasitoid 1 attacks the egg stage while parasitoid 2 attacks the larval or pupal stage. Second, the assumption can also apply when the parasitoids attack the same stage of the host but exhibit a competitive hierarchy. For instance, Chow and Mackauer [3] studied multiple parasitism of the pea aphid by the solitary hymenopterous parasites *Aphidius smithi* and *Praon pequodorum* in the laboratory. They found in larval competition, *P. pequodorum* was intrinsically superior to *A. smithi*, regardless of the latter's age. Finally, we assume that, on average, θ_i parasitoids emerge from a host parasitized by species i. Under these assumptions, the model is given by

$$(2.1)\qquad \begin{cases} N' &= \lambda f(N) N g_1(E_1) g_2(E_2)\cdots g_n(E_n),\\ P_1' &= \theta_1 f(N) N\{1-g_1(E_1)\},\\ P_2' &= \theta_2 f(N) N g_1(E_1)\{1-g_2(E_2)\},\\ &\vdots\\ P_n' &= \theta_n f(N) N g_1(E_1)\cdots g_{n-1}(E_{n-1})\{1-g_n(E_n)\}, \end{cases}$$

where N' and P_i' are the densities of the host and parasitoids, respectively, in the next generation. The state space for the host-parasitoid dynamics is $\mathbf{R}_+^{n+1} = \{(N,P) \in \mathbf{R}\times\mathbf{R}^n : N \ge 0, P_i \ge 0 \text{ for all } i\}$.

To complete the model, it is necessary to specify the density-dependent survivorship function $f(N)$, the encounter rate function E, and the escape functions g_i. Throughout this article, we assume the following:

A1. f is a continuous decreasing positive function such that $f(0) = 1$ and $\lim_{N\to\infty} f(N) = 0$.

Survivorship functions that satisfy assumption A1 include the generalized Beverton–Holt function $f(N) = \frac{1}{1+\alpha N^\beta}$ with $\alpha > 0$ and $\beta > 0$, the Ricker function $f(N) = \exp(-\alpha N)$ with $\alpha > 0$, and the Hassell function $f(N) = \frac{1}{(1+\alpha N)^\beta}$. To simultaneously account for search limitation and egg limitation, we follow the approach of Rogers [26] and define the average host encounter rate as

$$E_i = \frac{\alpha_i P_i}{1+\alpha_i b_i N},$$

where α is the searching efficiency of the parasitoid and b_i corresponds to the handling time or egg limitation of the parasitoid. For parsimony, we rewrite this average encounter rate as

$$E = \frac{P_i}{a_i + b_i N}, \tag{2.2}$$

where $a_i = \frac{1}{\alpha_i}$. One can view a_i as a measurement of search limitation. When there is no egg limitation (i.e., $b_i = 0$), the encounter rate reduces to the classical Nicholson–Bailey search limited encounter rate of $E_i = P_i/a_i$. Alternatively, when there is no search limitation (i.e., $a_i = 0$), the encounter rate reduces to the Thompson model $E_i = P_i/(b_i N)$ of egg-limited encounter rates [27, 31]. If eggs are randomly laid on hosts, then the fraction of hosts escaping parasitism is $\exp(-E_i)$. More generally, the Poisson escape term $\exp(-E_i)$ can be viewed as a limiting case of the negative binomial escape term $(1+E_i/k_i)^{-k_i}$ as $k_i \uparrow \infty$. This negative binomial escape function is commonly used to model nonrandom or aggregated parasitism events [6, 9, 13, 20]. In particular, $1/k_i$ can be interpreted as the coefficient of variation squared (CV^2) of the host encounter rate [11]. Consequently, larger values of k_i correspond to parasitic attacks being more evenly distributed across the hosts, while smaller values of k_i correspond to parasitoid attacks being aggregated on fewer hosts. To allow for this continuum of possibilities, we assume the following:

A2. $g_i(E_i) = \left(1 + \frac{E_i}{k_i}\right)^{-k_i}$ and $E_i = \frac{P_i}{a_i + b_i N_i}$ with $k_i > 0$ (possibly ∞), $a_i \geq 0$, and $b_i \geq 0$.

For ease of exposition, we write $k_i = \infty$ to refer to the Poisson escape function. The most important feature of escape function for the analysis is that $1/g_i$ is a concave function when $k_i < 1$ and $1/g_i$ is a convex function when $k_i > 1$.

Finally, to keep things meaningful, we assume the following:

A3. $\lambda, \theta_1, \ldots, \theta_n > 0$.

3. Egg-limited dynamics. Throughout this section, we assume that $a_i = 0$; i.e., there is no search limitation. For this case, we can make the change of variables

$$x = N, y_1 = E_1 = \frac{P_1}{b_1 N}, \ldots, y_n = E_n = \frac{P_n}{b_n N},$$

for which the dynamics of (2.1) partially decouple as follows:

$$\left\{ \begin{array}{rcl} x' &=& \lambda f(x) x g_1(y_1) g_2(y_2) \cdots g_n(y_n), \\ y_1' &=& \frac{\theta_1}{b_1 \lambda} \left(\frac{1}{g_1(y_1)} - 1 \right) \frac{1}{g_2(y_2)} \frac{1}{g_3(y_3)} \cdots \frac{1}{g_n(y_n)}, \\ y_2' &=& \frac{\theta_2}{b_2 \lambda} \left(\frac{1}{g_2(y_2)} - 1 \right) \frac{1}{g_3(y_3)} \cdots \frac{1}{g_n(y_n)}, \\ & \vdots & \\ y_{n-1}' &=& \frac{\theta_{n-1}}{b_{n-1} \lambda} \left(\frac{1}{g_{n-1}(y_{n-1})} - 1 \right) \frac{1}{g_n(y_n)}, \\ y_n' &=& \frac{\theta_n}{b_n \lambda} \left(\frac{1}{g_n(y_n)} - 1 \right). \end{array} \right. \tag{3.1}$$

To state our main result for this system, we need the following definition. Note that each $\frac{1}{g_i(y_i)} - 1$ is an increasing and strictly convex or concave function through the origin under assumption A2 and $k_i \neq 1$. Consequently, the nonnegative y_i^* defined below exist.

DEFINITION 3.1. *Assume* A2, A3, *and $k_i \neq 1$ for all i. Let $C = \{i : k_i > 1\}$. Define y_n^* to be the largest root of $y_n = \frac{\theta_n}{b_n \lambda}\left(\frac{1}{g_n(y_n)} - 1\right)$. Assuming y_j^* is defined for*

$j = i+1, \dots, n$, define y_i^* to be the largest root of

$$y_i = \frac{\theta_i}{b_i\lambda}\left(\frac{1}{g_i(y_i)} - 1\right)\frac{1}{g_{i+1}(\hat{y}_{i+1})}\cdots\frac{1}{g_n(\hat{y}_n)},$$

where

$$\hat{y}_j = \begin{cases} 0 & if\ j \in C, \\ y_j^* & if\ j \notin C. \end{cases}$$

Our main result is the following theorem. A key quantity in this theorem is $\lambda g_1(y_1)\dots g_n(y_n)$, which corresponds to the expected number of progeny produced per host.

THEOREM 3.2. *Assume* A1–A3, *$k_i \neq 1$ for all i, $a_i = 0$ for all i, $\theta_i \neq b_i\lambda\prod_{j=i+1}^n g_j(\hat{y}_j)$ for $1 \leq i \leq n-1$, and $\theta_n \neq b_n\lambda$. Let $C = \{i : k_i > 1\}$. Then we have the following:*

Host extinction. *If $y_i^* = 0$ for some $i \in C$, or $y_i^* > 0$ for all $i \in C$ (possibly $C = \emptyset$) and $\lambda g_1(\hat{y}_1)\dots g_n(\hat{y}_n) < 1$, then*

$$\lim_{t\to\infty}(N(t), P_1(t), \dots, P_n(t)) = (0, 0, \dots, 0)$$

whenever $N(0)\prod_{i=1}^n P_i(0) > 0$.

Host persistence. *If $C = \emptyset$ and $\lambda g_1(y_1^*)\dots g_n(y_n^*) > 1$, then there exists a positive constant $\delta > 0$ such that*

$$\liminf_{t\to\infty} N(t) \geq \delta \text{ and } \lim_{t\to\infty}\frac{P_i(t)}{b_i N(t)} = y_i^*$$

for all i whenever $N(0)\prod_{i=1}^n P_i(0) > 0$.

Conditional extinction. *If $y_i^* > 0$ for all $i \in C$, $C \neq \emptyset$, and $\lambda g_1(\hat{y}_1)\dots g_n(\hat{y}_n) > 1$, then there exist Borel sets $U, V \subset \mathbf{R}_+^{n+1}$ and $\delta > 0$ such that*

$$\liminf_{t\to\infty} N(t) \geq \delta, \qquad \lim_{t\to\infty}\prod_{i=1}^n P_i(t) = 0$$

whenever $(N(0), P_1(0), \dots, P_n(0)) \in U$, and

$$\lim_{t\to\infty}(N(t), P_1(t), \dots, P_n(t)) = (0, 0, \dots, 0)$$

whenever $(N(0), P_1(0), \dots, P_n(0)) \in V$. Moreover, U has positive (possibly infinite) Lebesgue measure, V has infinite Lebesgue measure, and $\mathbf{R}_+^{n+1} \setminus (U \cup V)$ has Lebesgue measure zero.

Theorem 3.2 (modulo equalities) characterizes the persistence and extinction dynamics of (2.1). In particular, host extinction can occur in two ways. If the host intrinsic fitness λ is less than one, then the host is unable to sustain itself and becomes extinct. Alternatively if $\lambda > 1$, then the host can persist in the absence of the parasitoids. However, if either $k_i > 1$ and $y_i^* = 0$ for a parasitoid or $\lambda\prod_{i=1}^n g_i(\hat{y}_i) < 1$, then the parasitoids drive the host extinct. Unconditional persistence of the host can occur only if the parasitoid attacks are sufficiently aggregated (i.e., $k_i < 1$ for all i) and the parasitoids do not overexploit their host (i.e., $\lambda\prod_{i=1}^n g_i(y_i^*) > 1$). How these different outcomes depend on the degree of egg limitation is illustrated in Figure 3.1

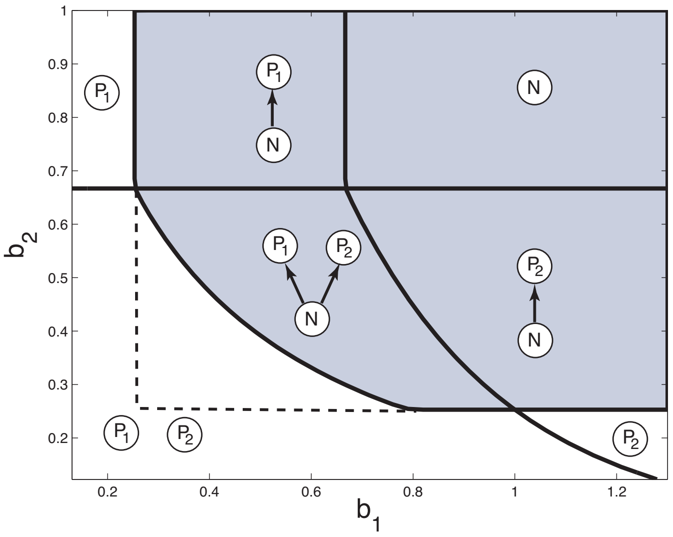


FIG. 3.1. *Ecological outcomes and how they vary with egg limitation. There are two parasitoids with conversion efficiencies $\theta_i = 1$ and aggregation parameters $k_i = 0.2$. The intrinsic fitness of the host is $\lambda = 1.5$. In the shaded region, the species that persist are shown. In the unshaded region, the host is driven to extinction by the indicated parasitoid(s). The dashed line delineates the region where both parasitoids but not a single parasitoid can drive the host to extinction.*

for parasitoids whose attacks are sufficiently aggregated. When egg limitation is sufficiently severe (i.e., b_i is sufficiently large) for a parasitoid species, the parasitoid is unable to establish itself. When egg limitation is sufficiently weak for a parasitoid species, it drives the host extinct. At intermediate levels of egg limitation, multiple parasitoids can drive the host to extinction when a single parasitoid species cannot (Figure 3.2 with parameter values from the dashed region in Figure 3.1).

When parasitoid attacks are sufficiently aggregated (i.e., $k_i < 1$ for all i), our results imply that the host dynamics have the limiting equation

$$N' = \lambda f(N) N \prod_{i=1}^{n} g_i(y_i^*),$$

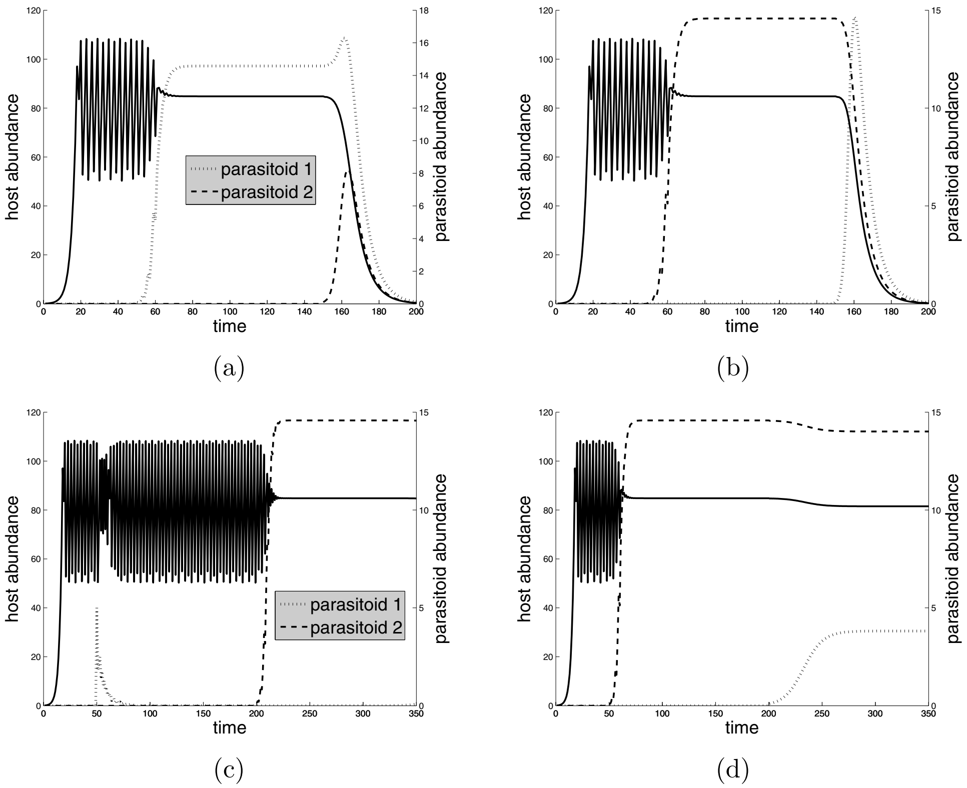
and the parasitoid dynamics track the host dynamics; i.e., asymptotically the ratio of the parasitoid to the host approaches $b_i y_i^*$ for parasitoid species i. Consequently, in this case a lot more can be said about the dynamics provided that the dynamics of the host are well understood. For instance, when the host dynamics can be described by the Beverton–Holt model, we get the following corollary of Theorem 3.2.

COROLLARY 3.3. *Suppose that $f(N) = \frac{1}{1+\alpha N}$, $k_i < 1$ for all i, and $\lambda^* := \lambda \prod_{i=1}^{n} g_i(y_i^*) > 1$. Then*

$$\lim_{t\to\infty} N(t) = \frac{\lambda^* - 1}{\alpha}, \qquad \lim_{t\to\infty} P_i(t) = \frac{b_i y_i^*(\lambda^* - 1)}{\alpha}$$

whenever $N(0)\prod_{i=1}^{n} P_i(0) > 0$.

Proof. Since $k_i < 1$ for all i and $\lambda^* > 1$, the second assertion of Theorem 3.2 applies. Let $Z(t) = (N(t), P_1(t), \dots, P_n(t))$ be a solution with $N(0)\prod_{i=1}^{n} P_i(0) > 0$.



(a) (b) (c) (d)

Fig. 3.2. *The effect of introducing one parasitoid at a time on the host-parasitoid dynamics. In* (a) *and* (c), *parasitoid* 1 *is introduced first, while parasitoid* 2 *is introduced later. In* (b) *and* (d), *the introduction order of parasitoids is reversed. In* (a) *and* (b), *the host is driven extinct only when both parasitoids are present. In* (c) *and* (d), *parasitoid* 1 *can establish itself only after parasitoid* 2 *has been established. In all figures,* $\lambda = 1.5$, $k_i = 0.2$, $f(N) = 1/(1 + (0.01N)^{10})$, *and* $\theta_i = 1.0$. *In* (a) *and* (b), $b_1 = b_2 = 0.4$. *In* (c) *and* (d), $b_1 = 0.75$ *and* $b_2 = 0.4$.

A result of Robinson [25] implies that the ω-limit set of $Z(t)$ is a chain recurrent set (see [25] for a definition). Theorem 3.2 implies that $\lim_{t\to\infty} \frac{P_i(t)}{N(t)} = b_i y_i^*$ for all i and $\liminf_{t\to\infty} N(t) > 0$. Since the only chain recurrent set in the invariant ray $\{(\eta, b_1 y_1^* \eta, \dots, b_n y_n^* \eta) : \eta > 0\}$ is the equilibrium $\frac{\lambda^* - 1}{\alpha}(1, y_1^* b_1, \dots, y_n^* b_n)$, the corollary follows. □

4. Proof of Theorem 3.2.

We begin with a lemma that shows that (2.1) is dissipative.

Lemma 4.1. *Assume* A1–A3. *There exists a constant* $M > 0$ *such that*

$$\limsup_{t\to\infty} N(t) \le M, \qquad \limsup_{t\to\infty} P_i(t) \le M$$

for all solutions $(N(t), P_1(t), \dots, P_n(t))$ *to* (2.1).

Proof. Assumption A1 implies that there exists $M_1 > 0$ such that $\lambda f(x) < 0.9$ for all $x \ge M_1$. Define $M_2 = \max\{M_1, \lambda M_1\}$ and

$$M = \max\{M_2, \theta_1 M_2, \dots, \theta_n M_2\}.$$

Let $(N(t), P_1(t), \dots, P_n(t))$ be a solution to (2.1). First, we will show that there exists a $T \geq 0$ such that $N(T) \leq M_2$. If $N(0) \leq M_2$, then we are done. Suppose $N(0), \dots, N(t) > M_2$. Since $g_1 \leq 1, \dots, g_n \leq 1$, f is decreasing, and $M_2 \geq M_1$, it follows that

$$N(t) = \lambda f(N(t-1))N(t-1)g_1(y_1(t-1)) \dots g_n(y_n(t-1)) \leq 0.9N(t-1).$$

Induction implies that $N(t) \leq 0.9^t N(0)$. Therefore, there exists $T \geq 0$ such that $N(T) \leq M_2$. Next, suppose that $N(T), \dots, N(T+t) \leq M_2$. Since $g_1 \leq 1, \dots, g_n \leq 1$, and f is decreasing, $N(T+t+1) \leq 0.9N(T+t) \leq 0.9M_2$ if $N(T+t) \geq M_1$, else $N(T+t+1) \leq \lambda M_1 \leq M_2$. Hence, induction implies that $N(t) \leq M_2 \leq M$ for all $t \geq T$. Finally, since $f \leq 1$ and $g_i \in [0,1]$ for $1 \leq i \leq n$,

$$P_i(t+1) \leq \theta_i N(t) \leq \theta_i M_2 \leq M$$

for all $t \geq T$. □

Define

$$G_i(y_i) = \frac{\theta_i}{b_i \lambda}\left(\frac{1}{g_i(y_i)} - 1\right).$$

LEMMA 4.2. *Assume* A2, A3, *and* $k_i \neq 1$. *Then*

$$z_i = c\, G_i(z_i) \tag{4.1}$$

has a nonnegative root for every $c \geq 0$. *For every* $c \geq 0$ *define* $z_i^*(c)$ *by the largest root of* (4.1). *Then the function* $z_i^* : \mathbf{R}_+ \to \mathbf{R}_+$ *is continuous.*

Proof. Since $0 = c\,G_i(0)$, $z_i = 0$ is always a root of (4.1). The function G_i is increasing and either strictly concave or strictly convex. Therefore, (4.1) has at most one positive root. This fact implies that z_i^* is a nonnegative function of c.

Consider the case where the function G_i is strictly concave, i.e., $k_i < 1$. In this case, (4.1) has a unique positive root if and only if $c > 1/G_i'(0)$. Therefore, $z_i^*(c) = 0$ if $c \in [0, 1/G_i'(0)]$ and $z_i^*(c) > 0$ if $c \in (1/G_i'(0), \infty)$. Since $z_i^*(c)$ is clearly continuous in $[0, 1/G_i'(0))$, it remains to show its continuity in $[1/G_i'(0), \infty)$. Define

$$F(c, z_i) = \begin{cases} 1 - cG_i'(0) & \text{if } z_i = 0, \\ 1 - cG_i(z_i)/z_i & \text{if } z_i > 0. \end{cases}$$

Then, by definition, $F(c, z_i^*(c)) = 0$ for all $c \in [1/G_i'(0), \infty)$. Furthermore, we can show that for each $c \in [1/G_i'(0), \infty)$,

$$\frac{\partial F}{\partial z_i} = \begin{cases} \frac{c\theta_i(1-k_i)}{2\lambda b_i k_i} \neq 0 & \text{if } z_i = 0, \\ \frac{c\theta_i}{\lambda b_i z_i^2}\left\{\left(1 + \frac{1-k_i}{k_i} z_i\right)\left(1 + \frac{z_i}{k_i}\right)^{k_i - 1} - 1\right\} \neq 0 & \text{if } z_i > 0. \end{cases}$$

Thus the application of the implicit function theorem to $F(c, z) = 0$ at $(c^*, z_i(c^*))$ with $c^* \in [1/G_i'(0), \infty)$ implies that there exists a continuous function $h(c)$ such that $F(c, h(c)) = 0$ holds in a neighborhood of $(c^*, z_i^*(c^*))$. Since a positive root of (4.1) is unique, h and z_i^* must be identical. The arbitrariness of c^* implies that $z_i^*(c)$ is continuous.

The case $k_i > 1$ can be proved similarly. In this case, (4.1) has a unique positive root if and only if $c\,G_i'(0) < 1$. ▯

LEMMA 4.3. *Assume* A2–A3 *and* $k_i \neq 1$. *Let* $z_i^*(c)$ *be the same as in Lemma* 4.2, *and let* $y_i(t)$ *be a solution to*

$$y_i(t+1) = G_i(y_i(t))c(t),$$

where $c(t)$ *is a positive sequence with* $\liminf_{t\to\infty} c(t) = \underline{c}$ *and* $\limsup_{t\to\infty} c(t) = \overline{c}$. *If* $k_i < 1$, *then*

$$z_i^*(\overline{c}) \geq \limsup_{t\to\infty} y_i(t) \geq \liminf_{t\to\infty} y_i(t) \geq z_i^*(\underline{c}).$$

If $k_i > 1$ *and* $\overline{c} = \underline{c}$, *then either*

$$\lim_{t\to\infty} y_i(t) = \infty, \lim_{t\to\infty} y_i(t) = z_i^*(\underline{c}) \textit{ or } \lim_{t\to\infty} y_i(t) = 0.$$

Proof. Suppose that $k_i < 1$. By assumption, for each $\epsilon > 0$ there exists a $T \geq 0$ such that

$$G_i(y_i(t))(\overline{c} + \epsilon) \geq y_i(t+1) \geq G_i(y_i(t))(\underline{c} - \epsilon)$$

for all $t \geq T$. Let $z(t)$ and $w(t)$ be the solutions of

$$z(t+1) = G_i(z(t))(\underline{c} - \epsilon) \text{ and } w(t+1) = G_i(w(t))(\overline{c} + \epsilon)$$

with $z(T) = w(T) = y_i(T)$. Then, by the monotonicity of G_i, $w(t) \geq y_i(t) \geq z(t)$ holds for all $t \geq T$. Since G_i is concave and $\lim_{x\to\infty} G_i(x)/x = 0$, it follows that $\lim_{t\to\infty} z(t) = z_i^*(\underline{c} - \epsilon)$ and $\lim_{t\to\infty} w(t) = z_i^*(\overline{c} + \epsilon)$. Therefore, we have

$$z_i^*(\overline{c} + \epsilon) \geq \limsup_{t\to\infty} y_i(t) \geq \liminf_{t\to\infty} y_i(t) \geq z_i^*(\underline{c} - \epsilon).$$

Since $\epsilon > 0$ is arbitrary and $z_i^*(c)$ is a continuous function, this inequality implies the first statement of the lemma.

Suppose that $k_i > 1$ and $\underline{c} = \overline{c}$. Suppose that $\limsup_{t\to\infty} y_i(t) < \infty$. Then the limit set of $y_i(t)$ is a compact internally chain recurrent set (see, e.g., [1]) for the dynamics of $y_i' = G_i(y_i)\underline{c}$. Since the only internally chain recurrent sets are the equilibria, 0, and $z_i^*(\underline{c})$ (possibly also 0), $\lim_{t\to\infty} y_i(t) = 0$ or $\lim_{t\to\infty} y_i(t) = z_i^*(\underline{c})$. Suppose that $\limsup_{t\to\infty} y_i(t) = \infty$. Since $\lim_{x\to\infty} G_i(x)/x = \infty$ (as $k_i > 1$) and G_i is convex, there exists $T > 0$, $M > 0$, and $\epsilon > 0$ such that

$$G_i(x)c(t) \geq (1+\epsilon)x$$

for all $t \geq T$ and $x \geq M$. Choose $T_2 > T$ such that $y_i(T_2) \geq M$. Then $y_i(T_2 + t) \geq (1+\epsilon)^t y_i(T_2)$ for all $t \geq 0$. Hence, $\lim_{t\to\infty} y_i(t) = \infty$. ▯

Let $(x(t), y_1(t), \ldots, y_n(t))$ be a positive solution to (3.1). An important implication of Lemma 4.3 is that $\liminf_{t\to\infty} y_i(t) \geq \hat{y}_i$. Indeed, Lemma 4.3 with $c(t) = 1$ applied to $y_n(t)$ implies $\liminf_{t\to\infty} y_n(t) \geq \hat{y}_n$. Suppose that $\liminf_{t\to\infty} y_i(t) \geq \hat{y}_i$ for $i = j+1, \ldots, n$. To prove the assertion for $i = j$, consider two cases. If $k_j > 1$, then $\hat{y}_j = 0$ and the assertion holds. If $k_j < 1$, then apply Lemma 4.3 with $c(t) = \prod_{i=j+1}^{n} 1/g_i(y_i(t))$.

To prove the first assertion of Theorem 3.2 about unconditional host extinction, we consider two cases. First, suppose that $y_i^* = 0$ for some $i \in C$. Then $G_i'(0)\hat{c} > 1$, where $\hat{c} = \frac{1}{g_{i+1}(\hat{y}_{i+1})} \cdots \frac{1}{g_n(\hat{y}_n)}$. Since $\liminf_{t\to\infty} y_j(t) \geq \hat{y}_j$ for all $i+1 \leq j \leq n$, continuity and monotonicity of g_j for $i \leq j \leq n$ imply that there exists a $T \geq 0$ and an $\eta > 1$ such that $y_i(t+1) \geq \eta y_i(t)$ for all $t \geq T$. Hence, $\lim_{t\to\infty} y_i(t) = \infty$. Since $P_i(t)$ is bounded by Lemma 4.1, it follows that $\lim_{t\to\infty} x(t) = 0$. For the second case, we assume that $y_i^* > 0$ for all $i \in C$ and $\lambda f(0) g_1(\hat{y}_1) \cdots g_n(\hat{y}_n) < 1$. Since $\liminf_{t\to\infty} y_i(t) \geq \hat{y}_i$ for all i and g_i are decreasing functions, there exist constants $\lambda_M < 1$ and $T \geq 0$ such that $\lambda f(0) g_1(y_1(t)) \cdots g_n(y_n(t)) \leq \lambda_M$ for all $t \geq T$. Therefore, $x(t+1) \leq \lambda_M x(t)$ holds for all $t \geq T$. Hence $\lim_{t\to\infty} x(t) = 0$.

To prove the second assertion of Theorem 3.2 about unconditional host persistence, assume that $C = \emptyset$ and $\lambda f(0) g_1(y_1^*) \cdots g_n(y_n^*) > 1$. Applying Lemma 4.3 inductively to $y_i(t)$ with $c(t) = \prod_{j=i+1}^n 1/g_j(y_j(t))$ implies that $\lim_{t\to\infty} y_i(t) = \hat{y}_i = y_i^*$ for all i. By the continuity of g_i, there exist $\lambda_M \geq \lambda_m > 1$ and $T_1 \geq 0$ such that

$$\lambda g_1(y_1(t)) \cdots g_n(y_n(t)) \in [\lambda_m, \lambda_M]$$

for all $t \geq T_1$. Since f is continuous, we can choose $\delta > 0$ such that $\lambda_m f(x) > 1$ for $x \in [0, \delta]$. Define $\alpha = \inf\{\lambda_m f(x) x : x > \delta\}$. Suppose $\alpha > 0$. Let $m = \min\{\delta, \alpha\}$. Since $\lambda_m f(x) x > x$ for all $x \in (0, m)$, there exists $T_2 \geq T_1$ such that $x(T_2) \in [m, \infty)$. By the definition of m, x does not escape from the interval (m, ∞). Finally, suppose $\alpha = 0$. Since $\lambda_M f(x) > 1$ for all $x \in [0, \delta]$ and $\lambda_M f(x) < 1$ for some $x > \delta$, the continuity of f ensures that the equation $\lambda_M f(x) = 1$ has a positive solution in the interval (δ, ∞). Since f is decreasing, there exists a unique positive solution, say $\bar{x} \in (\delta, \infty)$. Let $M = \max\{\lambda_M f(x) x : x \in [0, \bar{x}]\}$ and $\beta = \min\{\lambda_m f(x) x : x \in [\delta, M]\}$. Define $m = \min\{\delta, \beta\}$. By the definitions of m and M, if $x(t) \in [m, M]$ for some $t \geq T_1$, then $x(t)$ does not escape from the interval $[m, M]$. If $x(T_1) \in (M, \infty)$, then either $x(t) \in (M, \infty)$ for all $t \geq T_1$ or $x(T_2) \in (0, M]$ for some $T_2 \geq T_1$. Since the former case provides the desired conclusion, we consider the latter case. On the interval $(0, m)$, $\lambda_m f(x) x > x$ holds. Therefore, there exists a $T_3 \geq T_2$ such that $x(T_3) \in [m, M]$. This completes the proof of the second assertion of the theorem.

To prove the final assertion of Theorem 3.2, assume that $y_i^* > 0$ for all $i \in C$, $C \neq \emptyset$, and $\lambda g_1(\hat{y}_1) \ldots g_n(\hat{y}_n) > 1$. Assume $C = \{i_1, \ldots, i_k\}$ with $i_1 > i_2 > \cdots > i_k$. For each $1 \leq j \leq n$, define $U(j)$ as the set of initial conditions $(N(0), P_1(0), \ldots, P_n(0)) \in \mathbf{R}_+^{n+1}$ such that

$$N(0) \prod_{i=1}^n P_i(0) > 0 \text{ and } \lim_{t\to\infty} y_i(t) = \hat{y}_i \text{ for all } i \geq j,$$

and define $V(j)$ as the set of initial conditions such that

$$N(0) \prod_{i=1}^n P_i(0) > 0 \text{ and } \lim_{t\to\infty} y_i(t) = \infty \text{ for some } i \geq j.$$

For $j = i_1, i_2, \ldots, i_k$, we will prove inductively that $\mathbf{R}_+^{n+1} \backslash (U(j) \cup V(j))$ has Lebesgue measure zero.

As the first step of the induction, let $j = i_1$. If $j = n$, then convexity of G_n implies that $\lim_{t\to\infty} y_n(t) = \infty$ whenever $y_n(0) > y_n^*$, $y_n(t) = y_n^*$ for all t whenever $y_n(0) = y_n^*$, and $\lim_{t\to\infty} y_n(t) = 0$ whenever $y_n(0) < y_n^*$. Hence, $\mathbf{R}^{n+1} \backslash (U(j) \cup V(j))$ has Lebesgue measure zero. Assume that $j < n$. Since $k_i < 1$ for $i > j$, applying Lemma 4.3

inductively to $i = n, n-1, \dots, j+1$ with $c(t) = 1, 1/g_n(y_n(t)), \dots, \prod_{i=j+2}^{n} 1/g_i(y_i(t))$ implies that $\lim_{t\to\infty} y_i(t) = y_i^*$ for $i > j$ whenever $y_i(0) > 0$ for $i > j$. Applying Lemma 4.3 to $i = j$ with $c(t) = \prod_{i>j} 1/g_i(y_i(t))$ implies that either

$$\lim_{t\to\infty} y_j(t) = \hat{y}_j = 0, \ \lim_{t\to\infty} y_j(t) = y_j^*, \text{ or } \lim_{t\to\infty} y_j(t) = \infty.$$

The derivative of $y_j', \dots, y_n'$ in (3.1) with respect to $y_j, \dots, y_n$ is an upper triangular matrix whose diagonal elements are given by $d_i(y_i, \dots, y_n) = G_i'(y_i) \prod_{l=i+1}^{n} 1/g_l(y_l)$ for $i = j, \dots, n$. Since G_i are concave for $i = j+1, \dots, n$, G_j is convex, and $\theta_i \neq b_i \lambda \prod_{l=i+1}^{n} g_l(\hat{y}_l)$ for $1 \le i \le n$, it follows that $d_i(\hat{y}_i, \dots, \hat{y}_n) < 1$ for $i > j$ and $d_j(y_j^*, \hat{y}_{j+1}, \dots, \hat{y}_n) > 1$. Hence, $(y_j^*, \hat{y}_{j+1}, \dots, \hat{y}_n)$ is a hyperbolic equilibrium for the dynamics of (3.1) restricted to the $y_j, \dots, y_n$ subsystem. Moreover, the stable manifold of this equilibrium has codimension one in the $y_j, \dots, y_n$ hyperplane. Since the local stable manifold has Lebesque measure zero, and $y_j', \dots, y_n'$ in (3.1) is a diffeomorphism, the global stable manifold which is a countable union of preimages of the local stable manifold also has Lebesgue measure zero. Thus, $\mathbf{R}_+^{n+1} \setminus (U(j) \cup V(j))$ has Lebesgue measure zero.

For the next step of the induction, assume that $\mathbf{R}_+^{n+1} \setminus (U(j) \cup V(j))$ has Lebesgue measure zero for $j = i_1, \dots, i_l$. Let $j = i_{l+1}$. Suppose that $\lim_{t\to\infty} y_i(t) = \hat{y}_i$ for all $i \ge i_l$, i.e., $(N(0), P_1(0), \dots, P_n(0)) \in U(i_l)$. Since $k_i < 1$ for $i_{l+1} < i < i_l$, Lemma 4.3 applied inductively implies that $\lim_{t\to\infty} y_i(t) = \hat{y}_i$ for $i_{l+1} < i < i_l$, and consequently, $\lim_{t\to\infty} y_i(t) = \hat{y}_i$ for $i > i_{l+1}$. Since $k_j > 1$, Lemma 4.3 implies that either

$$\lim_{t\to\infty} y_j(t) = 0, \ \lim_{t\to\infty} y_j(t) = y_j^*, \text{ or } \lim_{t\to\infty} y_j(t) = \infty$$

Using an argument similar to the first step of the induction, the equilibrium $(y_j^*, \hat{y}_{j+1}, \dots, \hat{y}_n)$ is a hyperbolic equilibrium for the dynamics of (3.1) restricted to the $y_j, \dots, y_n$ hyperplane. Moreover, the stable manifold of this equilibrium has codimension greater than or equal to one. Hence, $U(i_l) \setminus (U(j) \cup V(j))$ has Lebesgue measure zero. Since $V(i_l) \subset V(j)$, $\mathbf{R}_+^{n+1} \setminus (U(j) \cup V(j))$ has Lebesgue measure zero.

Next, we need to show that $\mathbf{R}_+^{n+1} \setminus U(1) \cup V(1)$ has Lebesgue measure zero. If $i_k = 1$, then we are done by the prior induction. Assume that $i_k > 1$ and let $j = i_k$. Suppose that $\lim_{t\to\infty} y_i(t) = \hat{y}_i$ for $i \ge i_k$, i.e., $(N(0), P_1(0), \dots, P_n(0)) \in U(i_k)$. Applying Lemma 4.3 inductively implies that $\lim_{t\to\infty} y_i(t) = \hat{y}_i$ for all $i \ge 1$. Hence, $U(1) = U(i_k)$ and $V(1) = V(i_k)$. Thus, $\mathbf{R}_+^{n+1} \setminus (U(1) \cup V(1))$ has Lebesgue measure zero. Define $U = U(1)$ and $V = V(1)$. If $(N(0), P_1(0), \dots, P_n(0)) \in U$, then we can argue as in the proof of the second assertion of the theorem that $\liminf_{t\to\infty} N(t) \ge \delta$ for an appropriate choice of $\delta > 0$.

To complete the proof of the final assertion of Theorem 3.2, we need to show that $U(1)$ has positive (possibly infinite) Lebesque measure and that $V(1)$ has infinite Lebesgue measure. The equilibrium $(\hat{y}_1, \dots, \hat{y}_n)$ for $y_1', \dots, y_n'$ in (3.1) is linearly stable and its basin of attraction is an open subset of $\mathbf{R}_+^n$. Consequently, $U(1)$ is an open subset of $\mathbf{R}_+^{n+1}$ and has positive Lebesgue measure. To show that $V(1)$ has infinite Lebesgue measure, notice that if $y(0)$ is such that $|y_i(0) - y_i^*|$ is sufficiently small for $i > i_1$ (vacuously true if $i_1 = n$) and $y_{i_1}(0)$ is sufficiently large, then $\lim_{t\to\infty} y_{i_1}(t) = \infty$. Hence, $V(i_1)$ has infinite Lebesgue measure. Since $V(i_1) \subset V(1)$, the proof of the theorem is complete. □

5. Weakly search-limited parasitoids.

In this section, we examine the effect of including search limitation on the host-parasitoid dynamics.

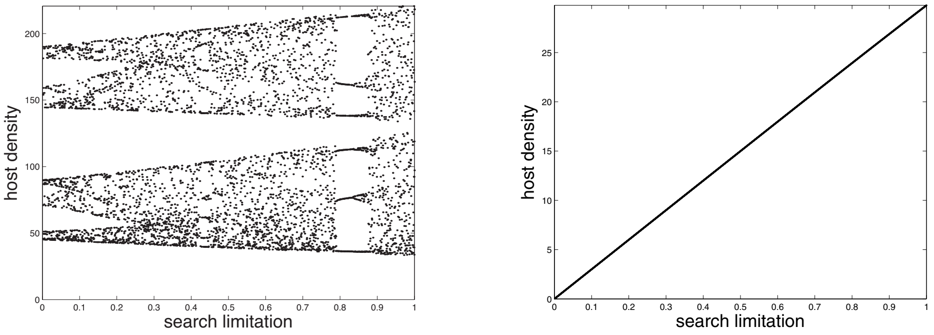
Proposition 5.1. *Assume* A1–A3. *If $\lambda > 1$ and $a_i > 0$, then there exists $\delta > 0$ such that*

$$\liminf_{t\to\infty} N(t) \geq \delta$$

for all solutions to (2.1) *with $N(0) > 0$.*

Proof. By dissipativity and continuity, system (2.1) has a compact forward invariant set that attracts all nonnegative solutions. Therefore, we can apply the theory of average Lyapunov functions (e.g., see Theorem 2.2 and Corollary 2.3 in [16]) and show that the face $N = 0$ is a repellor. More specifically, the application of the average Lyapunov function $L(N, P_1, \dots, P_n) = N$ shows that the face $N = 0$ is a repellor since every solution on the face $N = 0$ converges to the origin and $N'/N|_{(N,P_1,\dots,P_n)=(0,0,\dots,0)} = \lambda f(0)g_1(0)g_2(0)\cdots g_n(0) = \lambda > 1$. □

Although search-limited parasitoids cannot drive the host extinct, numerical simulations suggest that when purely egg-limited parasitoids can drive their host extinct, the inclusion of search limitation results in the host being suppressed to low equilibrium densities provided that parasitoid attacks are sufficiently aggregated (Figure 5.1b).



(a) only one parasitoid

(b) both parasitoids

Fig. 5.1. *The effect of single versus multiple parasitoid introductions when parasitoids exhibit search limitation. In* (a), *only one parasitoid is in the system. In* (b), *two parasitoids are in the system. In both figures, $\lambda = 10$, $k_i = 0.5$, $f(N) = 1/(1 + (0.01N)^4)$, $\theta_i = 2.0$, and $b_i = 0.1$. The parasitoids' searching limitations $a_1 = a_2$ vary as shown.*

Proposition 5.2. *Assume* A1–A3. *Let $(N(t), P_1(t), \dots, P_n(t))$ be a solution to* (2.1). *If $\lambda > 1$ and $\theta_m > b_m\lambda$, and $a_m > 0$ is sufficiently small, then there exists $\delta > 0$ such that*

$$\liminf_{t\to\infty} P_m(t) \geq \delta$$

whenever $N(0) > 0$ and $P_m(0) > 0$. Alternatively, if $\theta_i < b_i\lambda$ for $m \leq i \leq n$, then there exists a neighborhood U of the $P_m = P_{m+1} = \cdots = P_n = 0$ plane such that

$$\lim_{t\to\infty} (P_m(t) + \cdots + P_n(t)) = 0$$

whenever $(N(0), P_1(0), \dots, P_n(0))$ lies in U. Moreover, if $k_i < 1$ for $m \leq i \leq n$, then $U = \mathbf{R}_+^{n+1}$.

The first statement of Proposition 5.2 shows that if the reproductive ability of the parasitoid is higher than that of the host ($\theta_m/b_m > \lambda$) and there is weak search limitation (small $a_m > 0$), then the parasitoid coexists with its host. This type of coexistence is not always guaranteed if $a_m = 0$. However, if $a_m > 0$ is not small, numerical simulations suggest that the parasitoid may become extinct (see Figure 4 in [28]). The second statement of Proposition 5.2 considers the alternative situation ($\theta_m/b_m < \lambda$). In this case, the parasitoid whose population density is initially low becomes extinct irrespective of the intensity of search limitation. This attractivity result holds globally if the distribution of the parasitoid attack is aggregated ($k_m < 1$). However, this is not true if $k_m > 1$, since the system can have a positive fixed point or attractor. For instance, Figure 5.2 shows an example where a second positive fixed point bifurcates from the origin at $a_1 = 0$. As a_1 increases, this fixed point is stabilized and finally disappears due to a saddle node bifurcation. In this example, we also find a stable invariant loop in advance of the stabilization of the second fixed point (see Figure 5.2). Therefore, intermediate degrees of search limitation can produce a bistable system in which the initially rare parasitoid becomes extinct but the initially abundant parasitoid survives (see Figure 5.3).

Proof. Assume that $\lambda > 1$ and $\theta_m > b_m\lambda$. Proposition 5.1 implies that there is a compact attractor Γ such that the Γ does not intersect the $N = 0$ plane and such that the ω-limit set of $Z(t) = (N(t), P_1(t), \dots, P_n(t))$ lies in Γ whenever $N(0) > 0$. We will show that Γ intersected with the plane $P_m = 0$ is a repellor whenever $a_m > 0$ is sufficiently small. Let $Z(t) = (N(t), P_1(t), \dots, P_n(t))$ be a solution to (2.1) with $N(0) > 0$ and $P_m(0) = 0$. Since the ω-limit set of $Z(t)$ lies in Γ, we get that

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{1}{t} \ln\left(\frac{N(t)}{N(0)}\right) \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln\left(f(N(s))g_1(E_1(s)) \dots g_n(E_n(s))\lambda\right), \end{aligned}$$

where $E_i(s) = \frac{P_i(s)}{a_i + b_i N(s)}$. It follows that

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln\left(f(N(s))g_1(E_1(s)) \dots g_n(E_n(s))\right) = -\ln\lambda. \tag{5.1}$$

Let

$$G_i(N, P_1, \dots, P_n) = \frac{\partial P_i'}{\partial P_i}.$$

The Lyapunov exponent [24] corresponding to the P_m direction is given by

$$\begin{aligned} &\liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln G_m(Z(s)) \\ &\quad = \liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln\left(\theta_m N(s) f(N(s)) g_1(E_1(s)) \dots g_{m-1}(E_i(s)) \frac{1}{a_m + b_m N(s)}\right) \\ &\quad = \liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln\left(\theta_m N(s) \frac{1}{g_m(E_m(s))} \cdots \frac{1}{g_n(E_n(s))} \frac{1}{a_m + b_m N(s)} \frac{1}{\lambda}\right) \\ &\quad \geq \liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \ln\left(\theta_m N(s) \frac{1}{\lambda(a_m + b_m N(s))}\right), \end{aligned} \tag{5.2}$$

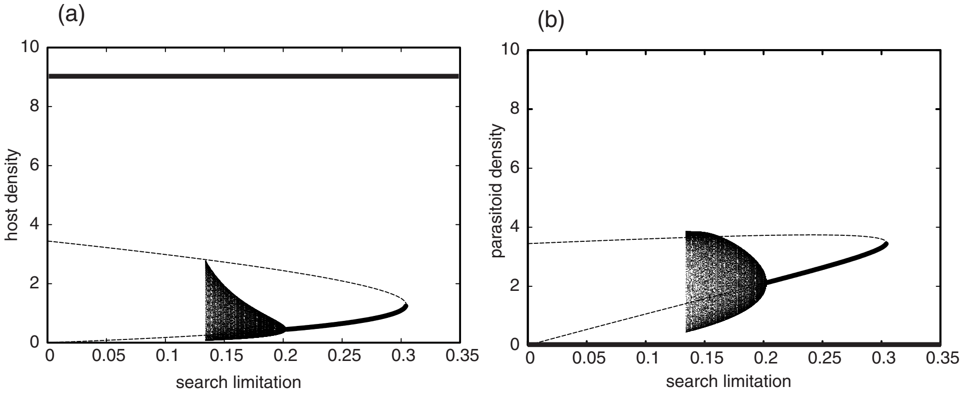


FIG. 5.2. *Bifurcation diagram of* (2.1) *with a single parasitoid species P_1, i.e., $n = 1$. The solid and dashed curves indicate stable and unstable fixed points, respectively. In both figures, $f(N) = 1/(1 + N)$, $\lambda = 10$, $b_1 = 1$, $k_1 = 2$, and $\theta_1 = 8$. If there is no search limitation ($a_1 = 0$), the parameters correspond to the conditional extinction case of Theorem* 3.2.

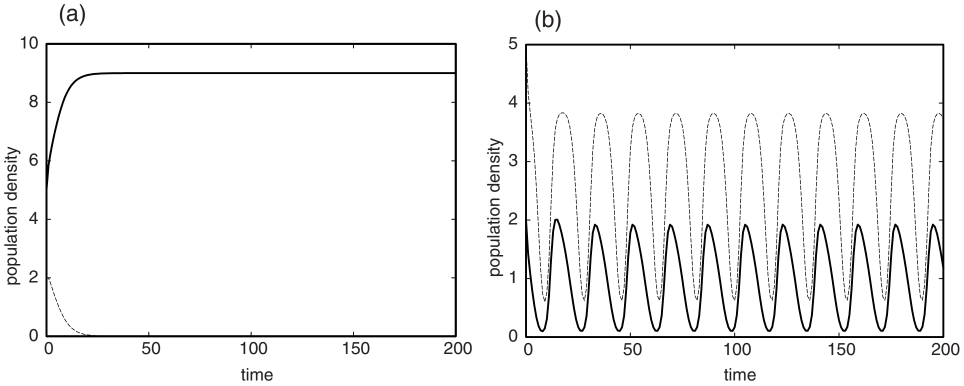


FIG. 5.3. *Bistable example of* (2.1) *with a single parasitoid species P_1, i.e., $n = 1$ The solid and dashed curves indicate the population density of host and parasitoid, respectively. In both figures, the survival function and parameters are $f(N) = 1/(1+N)$, $\lambda = 10$, $a_1 = 0.15$, $b_1 = 1$, $k_1 = 2$, and $\theta_1 = 8$. The initial conditions are* (a) $(N(0), P_1(0)) = (5, 2)$, *and* (b) $(N(0), P_1(0)) = (2, 5)$.

where the second line follows from (5.1) and the third line follows from $1/g_j \geq 1$ for all $1 \leq j \leq n$. If $\theta_m > b_m\lambda$ and $a_m > 0$ is sufficiently small, then (5.2) implies that there exists $\epsilon > 0$ such that

$$(5.3) \qquad \liminf_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \ln G_m(Z(s)) \geq \epsilon$$

for all initial conditions with $N(0) > 0$ and $P_m(0) = 0$. Equation (5.3) and the average Lyapunov theory (e.g., see Theorem 2.2 in [16]) with the average Lyapunov function $L(N, P_1, \dots, P_n) = P_m$ implies that Γ intersected with the $P_m = 0$ plane is a repellor. This completes the proof of the first statement.

Assume $\theta_i < b_i\lambda$ for $m \leq i \leq n$ and $a_i \geq 0$ for all i. If $\lambda < 1$, then $\lim_{t\to\infty} P_i(t) = 0$ for all i as $P_i(t+1) \leq \theta_i N(t)$ and we are done. Assume $\lambda > 1$. Proposition 5.1 implies that there is a compact attractor Γ such that Γ does not intersect the $N = 0$ plane and such that the ω-limit set of $Z(t) = (N(t), P_1(t), \dots, P_n(t))$ lies in Γ whenever $N(0) > 0$. Let Γ_1 be given by Γ intersected with the $P_m = P_{m+1} = \cdots = P_n = 0$ plane. We will show that Γ_1 is an attractor by proving that all the normal Lyapunov exponents (i.e., the Lyapunov exponents corresponding to Lyapunov directions that are not tangential to the $P_m = P_{m+1} = \cdots = P_n = 0$ plane) are negative (see, e.g., [35, Thm. 4] or [14]). To this end, consider a solution $Z(t) = (N(t), P_1(t), \dots, P_n(t))$ to (2.1) such that $Z(0) \in \Gamma_1$. Using (5.1) and assuming $m \le i \le n$, we get

$$\begin{aligned}
&\limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \ln G_i(Z(s))\\
&\quad= \limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \ln\left(\theta_i N(s)\frac{1}{\lambda}\frac{1}{g_i(E_i(s))}\cdots\frac{1}{g_n(E_n(s))}\frac{1}{a_i + b_i N(s)}\right)\\
&\quad= \limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \ln\left(\frac{\theta_i}{\lambda}\frac{N(s)}{a_i + b_i N(s)}\right)\\
&\quad\le \ln\left(\frac{\theta_i}{\lambda b_i}\right) < 0,
\end{aligned}$$

where the second line follows from $E_j(s) = 0$ for $m \le j \le n$ and the third line follows from $\frac{N}{a_i+b_iN} \le \frac{1}{b_i}$ whenever $N > 0$. Hence, all the normal Lyapunov exponents are negative and Γ_1 is an attractor. Next, assume that $k_i < 1$ for $m \le i \le n$. Let $Z(t)$ be a solution to (2.1) such that $N(0) > 0$. We will prove that

$$\lim_{t\to\infty} P_i(t) = 0 \text{ for } m \le i \le n$$

by induction on i. Consider $i = n$. If $P_n(0) = 0$, then we are done. Assume $P_n(0) > 0$. Then

$$\begin{aligned}
&\limsup_{t\to\infty} \frac{1}{t} \ln\left(\frac{P_n(t)}{P_n(0)}\right)\\
&\quad= \limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \ln\left(\theta_n N(s) f(N(s)) g_1(E_1(s)) \dots g_{n-1}(E_{n-1}(s)) \frac{(1 - g_n(E_n(s))}{P_n(s)}\right)\\
&\quad= \limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \ln\left(\theta_n N(s)\frac{1}{\lambda}\left(\frac{1}{g_n(E_n(s))} - 1\right)\frac{1}{P_n(s)}\right)\\
&\quad\le \limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \ln\left(\frac{\theta_n}{\lambda}\frac{N(s)}{a_n + b_n N(s)}\right)\\
&\quad\le \ln\left(\frac{\theta_n}{\lambda\, b_n}\right) < 0,
\end{aligned}$$

where the second line follows from (5.1), the third line follows from $1/g_n(x) - 1$ being concave (i.e., $k_n < 1$), and the fourth line follows from the fact that $\frac{N}{a_n+b_nN} \le \frac{1}{b_n}$ whenever $N > 0$. Hence, we have shown that $\lim_{t\to\infty} P_n(t) = 0$. Next, we proceed to the inductive step. Assume that $\lim_{t\to\infty} P_i(t) = 0$ for $j + 1 \le i \le n$, where $j \ge m$. If

$P_j(0)=0$, then we are done. Assume $P_j(0)>0$. Then

$$\begin{aligned}
&\limsup_{t\to\infty}\frac{1}{t}\ln\left(\frac{P_j(t)}{P_j(0)}\right)\\
&\quad=\limsup_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\ln\left(\theta_j N(s)f(N(s))g_1(E_1(s))\dots g_{j-1}(E_{j-1}(s))\frac{(1-g_j(E_j(s))}{P_j(s)}\right)\\
&\quad=\limsup_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\ln\left(\theta_j N(s)\frac{1}{\lambda}\frac{1}{g_{j+1}(E_{j+1}(s))}\cdots\frac{1}{g_n(E_n(s))}\left(\frac{1}{g_j(E_j(s))}-1\right)\frac{1}{P_j(s)}\right)\\
&\quad=\limsup_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\ln\left(\theta_j N(s)\frac{1}{\lambda}\left(\frac{1}{g_j(E_j(s))}-1\right)\frac{1}{P_j(s)}\right)\\
&\quad\le\limsup_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\ln\left(\frac{\theta_j}{\lambda}\frac{N(s)}{a_j+b_jN(s)}\right)\\
&\quad\le\ln\left(\frac{\theta_j}{\lambda b_j}\right)<0,
\end{aligned}$$

where the second line follows (5.1), the third line follows from induction (i.e., $g_i(0)=1$ for $j<i\le n$), the fourth line follows from $1/g_j(x)-1$ being concave (i.e., $k_j<1$), and the final line follows from the fact that $\frac{N}{a_j+b_jN}\le\frac{1}{b_j}$ whenever $N>0$. Hence, we have shown that $\lim_{t\to\infty}P_j(t)=0$, and the proof is complete. ☐

6. Discussion. We have studied the multiparasitoid-host dynamics described by (2.1). Under the assumption that each parasitoid is purely egg limited (i.e., not search limited), the dynamics of (2.1) have been classified sharply with respect to the extinction and persistence dynamics (see Theorem 3.2). Our main result implies that for the systems considered here, multiple parasitoids regulate a host population more efficiently than a single parasitoid. This conclusion can be derived for the following three scenarios in which the parasitoids can regulate the host:

(i) There are parasitoids with aggregated attacks ($k_i<1$ for all i) such that $\lambda g_1(y_1^*)\cdots g_n(y_n^*)<1$. This assembly of parasitoids drives the host extinct. The definition of y_i^* and concavity of $1/g_i$ (as $k_i<1$) imply that y_i^* is greater when you include more parasitoid species. Since $g_i(y_i^*)<1$ and g_i is a decreasing function for all i, the inequality $\lambda g_1(y_1^*)\cdots g_n(y_n^*)<1$ is more likely to hold if there are multiple aggregately distributed parasitoids irrespective of their superiority within a parasitized host (see Figures 3.1 and 3.2(a),(b)).

(ii) There are parasitoids with aggregated attacks ($k_i<1$ for all i) such that $\lambda g_1(\hat{y}_1)\cdots g_n(\hat{y}_n)>1$. This assembly of parasitoids does not drive the host extinct. Rather, they coexist with the host. As mentioned in section 3 (see also Corollary 3.3), after the establishment of these parasitoids, the dynamics of the host are asymptotic to

$$N'=\lambda f(N)N\prod_{i=1}^{n}g_i(y_i^*).$$

This equation suggests that introductions of multiple parasitoids lead to more efficient regulation of the host population. In fact, the last factor of the equation depresses the host density at a coexistence equilibrium.

(iii) There exists a highly reproductive parasitoid ($y_i^* = 0$) whose attacks are weakly aggregated ($k_i > 1$). By the definition of y_i^*, $y_i^* = 0$ if

$$\theta_i > b_i \lambda g_{i+1}(\hat{y}_{i+1}) \cdots g_n(\hat{y}_n).$$

This inequality is more likely to hold if there are multiple aggregately distributed parasitoids that are superior competitors within the parasitized host (i.e., P_j with $k_j < 1$, $j > i$).

The conclusion that multiple parasitoid introductions are more effective than single parasitoid introductions could depend on model assumptions. The main assumptions in our model are that the parasitoids are egg limited; the host suffers density-dependent mortality between the events of parasitism and death due to the parasitism; and the distributions of attacks of parasitoids are independent of each other. Kakehashi, Suzuki, and Iwasa [18] found that introductions disrupted host regulation when the distributions of parasitoid attacks completely overlap. Therefore, we expect that incorporation of overlapping parasitoid distributions into our model may lead to a similar conclusion. Interestingly, our conclusion seems insensitive to other assumptions. For example, our conclusion agrees with that of May and Hassell [21], who assume purely search-limited parasitism and no host density dependence.

REFERENCES

[1] M. Benaïm and M. W. Hirsch, *Asymptotic pseudotrajectories and chain recurrent flows, with applications*, J. Dynam. Differential Equations, 8 (1996), pp. 141–176.

[2] C. J. Briggs, *Competition among parasitoid species on a stage-structured host and its effect on host suppression*, The American Naturalist, 141 (1993), pp. 372–397.

[3] F. J. Chow and M. Mackauer, *Inter- and intraspecific larval competition in aphidius smithi and praon pequodorum (hymenoptera: aphidiidae)*, Canadian Entomologist, 116 (1984), pp. 1097–1107.

[4] P. DeBach, *Biological Control of Insect Pests and Weeds*, W. H. Rheinhold, New York, 1964.

[5] L. E. Ehler and R. W. Hall, *Evidence for competitive exclusion of introduced natural enemies in biological control*, Environmental Entomology, 11 (1982), pp. 1–4.

[6] W. M. Getz and N. J. Mills, *Host-parasitoid coexistence and egg-limited encounter rates*, The American Naturalist, 148 (1996), pp. 333–347.

[7] H. C. J. Godfray, *Parasitoids*, Princeton University Press, Princeton, NJ, 1994.

[8] D. J. Greathead and A. H. Greathead, *Biological control of insect pests by insect parasitoids and predators: The biocat database*, Biocontrol News Info., 13 (1992), pp. 61N–68N.

[9] M. P. Hassell, *The Dynamics of Arthropod Predator-Prey Systems*, Monographs in Population Biology 13, Princeton University Press, Princeton, NJ, 1978.

[10] M. P. Hassell and R. M. May, *Stability in insect host-parasite models*, Journal of Animal Ecology, 42 (1973), pp. 693–726.

[11] M. P. Hassell, R. M. May, S. W. Pacala, and P. L. Chesson, *The persistence of host-parasitoid associations in patchy environments.* I. *A general criterion*, The American Naturalist, 138 (1991), pp. 586–583.

[12] G. E. Heimpel and J. A. Rosenheim, *Egg limitation in parasitoids: A review of the evidence and a case study*, Biological Control, 11 (1998), pp. 160–168.

[13] M. E. Hochberg and R. D. Holt, *Refuge evolution and the population dynamics of coupled host-parasitoid associations*, Evolutionary Ecology, 9 (1995), pp. 633–661.

[14] J. Hofbauer and S. J. Schreiber, *To persist or not to persist?*, Nonlinearity, 17 (2004), pp. 1393–1406.

[15] C. B. Huffaker, *Biological Control*, Plenum, New York, 1971.

[16] V. Hutson, *A theorem on average Liapunov functions*, Monatsh. Math., 98 (1984), pp. 267–275.

[17] M. A. Jervis, G. E. Heimpel, P. N. Ferns, J. A. Harvey, and N. A. C. Kidd, *Life-history strategies in parasitoid wasps: A comparative analysis of 'ovigeny'*, Journal of Animal Ecology, 70 (2001), pp. 442–458.

[18] M. Kakehashi, Y. Suzuki, and Y. Iwasa, *Niche overlap of parasitoids in host-parasitoid systems: Its consequences to single versus multiple introduction controversy in biological control*, Journal of Applied Ecology, 21 (1984), pp. 115–131.

[19] S. D. Lane, N. J. Mills, and W. M. Getz, *The effects of parasitoid fecundity and host taxon on the biological control of insect pests: The relationship between theory and data*, Ecological Entomology, 24 (1999), pp. 181–190.

[20] R. M. May, *Host-parasitoid systems in patch environments: A phenomenological model*, Journal of Animal Ecology, 47 (1978), pp. 833–843.

[21] R. M. May and M. P. Hassell, *The dynamics of multiparasitoid-host interactions*, The American Naturalist, 117 (1981), pp. 234–261.

[22] R. M. May, M. P. Hassell, R. M. Anderson, and D. W. Tonkyn, *Density dependence in host-parasitoid models*, Journal of Animal Ecology, 50 (1981), pp. 855–865.

[23] N. J. Mills and W. M. Getz, *Modelling the biological control of insect pests: A review of host-parasitoid models*, Ecol. Model., 92 (1996), pp. 121–143.

[24] V. I. Oseledec, *A multiplicative ergodic theorem*, Trans. Moscow Math. Soc., 19 (1968), pp. 197–231.

[25] C. Robinson, *Stability theorems and hyperbolicity in dynamical systems*, Rocky Mountain J. Math., 7 (1977), pp. 425–437.

[26] D. J. Rogers, *Random search and insect population models*, Journal of Animal Ecology, 41 (1972), pp. 369–383.

[27] S. J. Schreiber, *Host-parasitoid dynamics of a generalized Thompson model*, J. Math. Biol., 52 (2006), pp. 719–732.

[28] S. J. Schreiber, *Periodicity, persistence, and collapse in host-parasitoid systems with egg limitation*, J. Biol. Dyn., 1 (2007), pp. 273–288.

[29] S. J. Schreiber, N. J. Mills, and A. P. Gutierrez, *Host-limited dynamics of autoparasitoids*, J. Theoret. Biol., 212 (2001), pp. 141–153.

[30] K. Shea, R. M. Nisbet, W. W. Murdoch, and H. J. S. Yoo, *The effect of egg limitation on stability in insect host-parasitoid population models*, Journal of Animal Ecology, 65 (1996), pp. 743–755.

[31] W. R. Thompson, *La theory mathematique de l'action des parasites entomophages et le facteur du hassard*, Ann. Fac. Sci. Marseille, 2 (1924), pp. 69–89.

[32] A. L. Turnbull and D. A. Chant, *The practice and theory of biological control of insects in Canada*, Canadian Journal of Zoology, 39 (1961), pp. 697–753.

[33] R. Van den Bosch, *Comments on population dynamics of exotic insects*, Bulletin of the Entomological Society of America, 14 (1968), pp. 112–115.

[34] K. E. F Watt, *Community stability and the strategy of biological control*, Canadian Entomologist, 97 (1965), pp. 887–895.

[35] L. Xu, B. Chen, Y. Zhao, and Y. Cao, *Normal Lyapunov exponents and asymptotically stable attractors*, Dyn. Syst., 23 (2008), pp. 207–218.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 977–998

SENSITIVITY AND ROBUSTNESS IN CHEMICAL REACTION NETWORKS[*]

GUY SHINAR[†], URI ALON[†], AND MARTIN FEINBERG[‡]

Abstract. For a wide class of chemical reaction networks, including all those governed by detailed balanced mass-action kinetics, we examine the robustness of equilibrium species concentrations against fluctuations in the overall reactant supply. In particular, we present lower bounds on the individual species-concentration sensitivities that derive from reaction network structure alone, independent of kinetic parameters or even of the particular equilibrium state at which sensitivities are calculated. These bounds suggest that, in the class of reaction networks considered here, very high robustness (i.e., very low sensitivities) should be expected only when the various molecules are constructed from a large number of distinct elemental building blocks that appear in high multiplicity or that combine gregariously. This situation is often encountered in biology.

Key words. equilibrium points, chemical reaction networks, sensitivity, robustness, systems biology

AMS subject classifications. 80A30, 37C05, 37C25, 92C45

DOI. 10.1137/080719820

1. Introduction.

1.1. Motivation.

Our interest is in understanding the relationship between the general features of a chemical reaction network and the sensitivity of its equilibria to changes in the overall supply of reactants.

Although our concerns are with reaction networks in general, we are motivated by questions that arise naturally in biology. The cell is highly dynamic. Nevertheless, it is reasonable to suppose that in the cytoplasm certain biochemical modules (reaction networks) act on a timescale that is fast relative to other cellular processes (e.g., relative to the production of large proteins or relative to the transport of small molecules across membranes). Thus, the concentrations of species participating in such a “fast” reaction network might be regarded as instantaneously at or near equilibrium.

At the same time, the “fast” biochemical module under consideration might be perturbed within the cell by slower changes in the overall reactant supply, as fresh proteins are produced or as smaller molecules enter the cell from its exterior. In response, the “fast” equilibrium composition would shift accordingly. For some purposes it might be advantageous for certain species concentrations within the module to be insensitive (robust) against such disturbances, while for other species great sensitivity might be highly desirable. Thus, as has been suggested by Veitia [25], it becomes important to understand how reaction network structure affects the responsiveness of its equilibrated species concentrations to changes in the ambient cellular environment.

[*] Received by the editors March 31, 2008; accepted for publication (in revised form) September 5, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/siap/69-4/71982.html

[†] Departments of Molecular Cell Biology and Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel (guy.shinar@weizmann.ac.il, uri.alon@weizmann.ac.il). The work of the first and second authors was supported by the Minerva Foundation, the United States National Institutes of Health, and the Kahn Family Foundation.

[‡] Department of Chemical Engineering and Department of Mathematics, Ohio State University, 125 Koffolt Laboratories, 140 W. 19th Avenue, Columbus, OH 43210 (feinberg.14@osu.edu). The work of the third author was supported by United States National Science Foundation grant BES-042549 and by an extended visit to the Department of Systems Biology, Harvard Medical School.

The present inquiry is not dissimilar in nature to other lines of research. One is biochemical systems theory pioneered by Savageau [23]. Another is metabolic control analysis developed from work of Higgins [15] by Kacser and Burns [19] and Heinrich and Rapoport [13, 14]. Here we provide general results along similar lines, but results that depend only on network structure. These results apply, within a large and relevant class of chemical reaction systems, to networks of arbitrarily large size and complexity.

For our purposes, it will suffice to presume that the reaction networks we study are governed by mass-action kinetics conforming to what chemists call *detailed balance* [26, 20, 10]. In fact, even these presumptions are overly strong, for we shall require only that the system considered have what Horn and Jackson called the *quasi-thermostatic* property [16]. (Detailed balancing in mass-action systems is sufficient but not necessary to ensure quasi-thermostatic behavior.)

The remainder of this introduction is organized as follows: In section 1.2 we define, in the context of an informal example, what we mean by *species sensitivities.* In section 1.3, we indicate the type of theory we seek by stating one principal result: lower bounds for species sensitivities *that derive from network structure alone.* In very rough terms, these bounds suggest that *strong robustness (i.e., very low sensitivities) of equilibrium concentrations against variations in reactant supply can be expected only when the species are built from highly gregarious or multiplicitous building blocks—that is, from certain* elements *that associate indiscriminately with each other or that appear in high multiplicity within the compound species.* In section 1.4 we discuss the organization of the remainder of the article.

1.2. Sensitivities. To illustrate more concretely the problems that are of interest, it will be useful to consider, informally, the relatively simple reaction network

$$2A \rightleftharpoons A_2, \qquad A_2 + B \rightleftharpoons A_2B, \qquad A_2B + C \rightleftharpoons A_2BC. \tag{1.1}$$

Note. Network (1.1) is inspired by a regulated recruitment [22] model for a gene transcription control mechanism: Species A corresponds to a protein monomer that can dimerize to form A_2, an active transcription factor. The dimer A_2 (but not the monomer A) can bind the DNA promoter B to form the compound A_2B. The compound A_2B can subsequently recruit the enzyme RNA polymerase, denoted C, to form the active compound A_2BC. It is the concentration of this last compound that determines the system's activity—the rate of transcription of the gene downstream from B.

Although we shall eventually consider a far wider class of reaction networks, our example here is an instance of what we shall later call a *constructive network*: Network (1.1) contains, explicitly, certain elemental species A, B, and C, from which all other species are ultimately constructed. (It needn't be the case in a constructive network that, as with our example, all reactions merely add a monomer of one of the elements. Thus, for example, a reaction such as $A_2B + A_2BC \rightleftharpoons A_4B_2C$ might also be included.) Certainly not all reaction networks are of this kind, but here and in section 1.3 consideration of constructive networks will facilitate both an introduction to the questions of interest and a description of some results.

As indicated earlier, we suppose that the individual reactions in network (1.1) are governed by mass-action kinetics, with fixed positive rate constants. (In the case of network (1.1), the detailed balance requirements are automatically satisfied, with no constraints imposed on rate constant values [10].) For the purposes of our discussion, we suppose also that the reactions take place in the context of a closed vessel, the contents of which are maintained at a fixed volume and temperature. Corresponding to a certain initial supply of various species, the mixture composition will evolve in time and ultimately approach equilibrium.

Note that the reactions conserve the elements A, B, and C, although these elements may reside latently within the compound species A_2, A_2B, and A_2BC. In fact, if $c_s(t)$ indicates the molar concentration of species s at time t, we expect that, for all time, the conservation conditions

$$\begin{aligned} &c_A(t) + 2c_{A_2}(t) + 2c_{A_2B}(t) + 2c_{A_2BC}(t) = T_A, \\ &c_{A_2B}(t) + c_B(t) + c_{A_2BC}(t) = T_B, \\ &c_C(t) + c_{A_2BC}(t) = T_C \end{aligned} \tag{1.2}$$

would be respected. In (1.2) we have denoted by T_A, T_B, and T_C the time-invariant *total* molar concentrations of the elements A, B, and C, regardless of whether they appear overtly or latently.

For specified positive values of the total element concentrations, say T_A^*, T_B^*, and T_C^*, the (polynomial) mass-action differential equations for network (1.1), formulated in the usual way [7], will give rise to precisely one equilibrium composition $c_A^*, c_B^*, c_C^*, c_{A_2}^*, c_{A_2B}^*, c_{A_2BC}^*$, consistent with that specification. On the other hand, if the vessel is temporarily opened and additional amounts of A, B, and C are added (either directly or by the addition of A_2, A_2B, and A_2BC), the resealed vessel will eventually come to a new and different equilibrium $c_A^{**}, c_B^{**}, c_C^{**}, c_{A_2}^{**}, c_{A_2B}^{**}, c_{A_2BC}^{**}$, which is the equilibrium consistent with the, now different, total concentrations T_A^{**}, T_B^{**}, and T_C^{**} of the elements. As we shall see later in the article, the equilibrium concentrations of the six species are given by smooth functions of T_A, T_B, and T_C.

Our interest is in the way that equilibrium concentrations of the various species are affected by small variations in the supply of the elements A, B, and C. By the *sensitivity matrix* for the system we mean the array whose entries are given by

$$\left(\frac{\partial \ln \bar{c}_s}{\partial \ln T_e}\right), \tag{1.3}$$

where $s \in \{A, B, C, A_2, A_2B, A_2BC\}$, $e \in \{A, B, C\}$, and $\bar{c}_s(\cdot,\cdot,\cdot)$ is the function that gives, for each specification of T_A, T_B, and T_C, the equilibrium concentration of species s. Values of the entries in the sensitivity matrix will depend, of course, on the equilibrium composition at which they are evaluated (in particular, on the values of T_A, T_B, and T_C at that equilibrium). The entry $\frac{\partial \ln \bar{c}_{A_2BC}}{\partial \ln T_B}$, for example, indicates the magnitude, at a particular equilibrium, of the fractional change in the equilibrium concentration of A_2BC in response to a small fractional change in the total concentration of element B. (The use of logarithms to reflect *fractional* changes, which is common in biology [23], is especially compelling when the various species concentrations can be of very different magnitudes.)

By the *sensitivity of species s* at equilibrium composition c^*, denoted $\Lambda^s(c^*)$, we mean the largest of the absolute values of the entries in the sensitivity matrix row corresponding to species s. Thus, for example,

$$\Lambda^{A_2BC}(c^*) = \max\left\{\left|\frac{\partial \ln \bar{c}_{A_2BC}}{\partial \ln T_A}\right|, \left|\frac{\partial \ln \bar{c}_{A_2BC}}{\partial \ln T_B}\right|, \left|\frac{\partial \ln \bar{c}_{A_2BC}}{\partial \ln T_C}\right|\right\}_{c^*}. \tag{1.4}$$

In this instance, $\Lambda^{A_2BC}(c^*)$ gives an indication, at equilibrium c^*, of the most pronounced of the variations in the equilibrium concentration of A_2BC, as the mixture is perturbed, respectively, by variations in the supply of A, B, and C.

Our interest is in providing means to calculate the sensitivity matrix at a particular equilibrium composition and, even more, in coming to a qualitative understanding of how reaction network character imposes *intrinsic* lower bounds on its species sensitivities, *bounds that derive from the reaction network alone, independent of rate constant values or even of the equilibrium composition at which they are calculated.* Bounds of this kind are described in the next section.

1.3. Network-imposed sensitivity bounds: Gregarious or multiplicitous elements are necessary for highly robust equilibria. Our goal in this section is to describe, informally and in the context of constructive networks, ways in which the network itself places a lower bound on species sensitivities. For a constructive network, we denote by $\mathscr{S}$ the set of species and by $\mathscr{E}$ the set of elements. In the case of network (1.1), $\mathscr{S} = \{A, B, C, A_2, A_2B, A_2BC\}$ and $\mathscr{E} = \{A, B, C\}$. For each element $e \in \mathscr{E}$ and each species $s \in \mathscr{S}$ we denote by M_e^s the *e-content* of species s—that is, the content of element e in species s. Thus, in our example, $M_A^A = 1$, $M_A^B = 0$, $M_A^C = 0$, $M_A^{A_2} = 2$, $M_A^{A_2B} = 2$, $M_A^{A_2BC} = 2$, and so on. (Later on, beginning in section 3, the notion of "element" will have meaning for reaction networks broadly, not just for those that are constructive. Even in this broader context it will make sense to speak of "the e-content of species s." Results described in this section will then extend to reaction networks in general.)

For each pair of elements $e, e' \in \mathscr{E}$ we denote by $M_e^{max}(e')$ the maximal e-content that can be found as we search over all species that contain the element e' (i.e., over all species that have positive e'-content). In network (1.1), the species that contain A are A, A_2, A_2B, and A_2BC. Thus, for example,

$$M_B^{max}(A) = \max\left\{M_B^A, M_B^{A_2}, M_B^{A_2B}, M_B^{A_2BC}\right\} = \max\{0, 0, 1, 1\} = 1.$$

Evaluating $M_e^{max}(e')$ for all combinations of $e, e' \in \mathscr{E}$ in network (1.1), we obtain

$$\begin{aligned} &M_A^{max}(A) = 2, \qquad M_A^{max}(B) = 2, \qquad M_A^{max}(C) = 2,\\ &M_B^{max}(A) = 1, \qquad M_B^{max}(B) = 1, \qquad M_B^{max}(C) = 1,\\ &M_C^{max}(A) = 1, \qquad M_C^{max}(B) = 1, \qquad M_C^{max}(C) = 1. \end{aligned} \tag{1.5}$$

Note that in general $M_e^{max}(e')$ will be high when there is an e'-containing species in which e appears with high multiplicity.

The *degree* of element e is given by

$$\deg(e) = \sum_{e' \in \mathscr{E}} M_e^{max}(e'). \tag{1.6}$$

Thus, for network (1.1),

$$\deg(A) = 6, \qquad \deg(B) = 3, \qquad \deg(C) = 3. \tag{1.7}$$

Note that a high degree for a particular element will result if it is free to combine with many other element partners (*resulting in many nonzero terms in* (1.6)) *or if it is multiplicitous in at least one instance of those various liaisons* (*resulting in a high* *value for at least one of the terms in* (1.6)). Indeed, a very high degree for a particular element *requires* such gregarious or multiplicitous couplings (or a combination of both).

All this is relevant to the description of a result we shall present as Theorem 7.3, a theorem that provides a *network-imposed* lower bound on the species sensitivities: *At every equilibrium composition c^* and for each species $s \in \mathscr{S}$,*

$$\Lambda^s(c^*) \geq \max_{e \in \mathscr{E}} \left\{ \frac{M_e^s}{\deg(e)} \right\}. \tag{1.8}$$

It should be noted that the lower bound afforded by (1.8) *is an attribute of the network alone. It is independent of rate constant values and of the particular equilibrium compositions at which sensitivities are of interest.* For network (1.1) we have at every equilibrium composition c^*,

$$\begin{aligned} &\Lambda^A(c^*) \geq \frac{1}{6}, \qquad \Lambda^B(c^*) \geq \frac{1}{3}, \qquad \Lambda^C(c^*) \geq \frac{1}{3}, \\ &\Lambda^{A_2}(c^*) \geq \frac{1}{3}, \qquad \Lambda^{A_2B}(c^*) \geq \frac{1}{3}, \qquad \Lambda^{A_2BC}(c^*) \geq \frac{1}{3}. \end{aligned} \tag{1.9}$$

The bound given by (1.8) *tells us that for an equilibrium concentration of a particular species s to be very robust against fluctuations in the overall supply of elements (that is, if s is to have very low sensitivity), the elements of which s is composed should have very high degree (so they should be gregarious or multiplicitous).* It helps too if species s itself is composed of very *few* copies of the various elements—that is, if M_e^s is low for the various $e \in \mathscr{E}$, especially those elements of low degree. In the setting of constructive systems, if we consider the sensitivities of the elements themselves (that is, for $s = e$), we note that $M_{e'}^e = 1$ when $e' = e$ and 0 otherwise. Thus, for each $e \in \mathscr{E}$, (1.8) reduces to

$$\Lambda^e(c^*) \geq \frac{1}{\deg(e)}. \tag{1.10}$$

As indicated earlier, we have stated here just one principal result, carried by (1.8), and then only in the context of what we have called constructive networks. That restriction was solely for the purpose of this introduction. Our concerns extend to networks in general, and our interests are not limited to the establishment of bounds. We seek, for example, to provide means to determine how entries in the sensitivity matrix depend on the current equilibrium state.

1.4. Organization.

In section 2 we provide an introduction to the rudimentary aspects of chemical reaction network theory, in particular to properties of *quasi-thermostatic* systems.

In section 3 we introduce the idea of *elemented reaction networks.* These constitute a broad generalization of what we call constructive reaction networks in this section: The fact is that the differential equations that derive from reaction networks typically reflect certain "conservation conditions" (integrals of motion), although what is conserved cannot always be clearly associated with total concentrations of species appearing overtly in the network (A, B, and C in our example). Nevertheless, one can generally associate with the network certain "elements," not necessarily overt species, whose total concentrations are conserved along solutions of the differential equations that the network induces. (An element might, for example, be identified with a moiety that manifests itself latently within the various species of the network while not itself appearing overtly.) For a given network there can be many such choices of (independently) conserved elements, and a particular application might favor one choice over another. By an elemented reaction network we mean a reaction network taken with one such choice of elements.

In section 4 we begin to examine properties of elemented quasi-thermostatic systems. This will set the stage for a discussion of sensitivities in section 5. Extended to elemented systems generally, our interest (as in this introductory section) will be in the variations of the equilibrium concentrations of the species in response to variations in the total concentrations of its elements. In section 6 we provide computational means to determine the sensitivity matrix—that is, to determine *through explicit formulas* how its entries depend on the particular equilibrium state at which they are calculated.

In section 7 we deduce lower bounds on the species sensitivities, bounds that are induced by the network alone, independent of rate constants and even of the equilibrium state at which the sensitivities are calculated. As in this introductory section, these bounds will be related to the *degrees* of the various elements. For elemented networks generally, the degree of an element is again influenced by the extent to which it combines in a gregarious and multiplicitous fashion. In section 8, we define formally what we mean by a constructive system, and we state a result that is particular to them. In section 9 we offer some brief concluding remarks.

2. Some ideas from chemical reaction network theory.

In this section we provide a brief review of rudimentary material from chemical reaction network theory [10, 16, 7, 5, 6, 8, 9, 2, 3, 4, 17, 11]. (An introduction for mathematicians, with more motivational discussion, can be found in [7].) Although chemical reaction network theory addresses a wide variety of dynamical issues, our focus here is exclusively on what Horn and Jackson [16] called *quasi-thermostatic* behavior.

2.1. Notation.

We denote the real numbers by $\mathbb{R}$, the strictly positive real numbers by $\mathbb{R}_+$, and the nonnegative real numbers by $\bar{\mathbb{R}}_+$. For an arbitrary finite set I (usually the set of species in a reaction network) we denote by $\mathbb{R}^I$ the real vector space of all formal sums $\sum_{i \in I} u_i i$ in which all u_i are real. By $\mathbb{R}^I_+$ (respectively, $\bar{\mathbb{R}}^I_+$) we mean the set of all $u \in \mathbb{R}^I$ for which all u_i are positive (respectively, nonnegative). By the *support* of an element $u \in \mathbb{R}^I$ we mean the subset of I defined by $\operatorname{supp}(u) = \{i \in I : u_i \neq 0\}$.

We use the symbol "$\circ$" to indicate componentwise multiplication. That is, for every u and v in $\mathbb{R}^I$, $u \circ v$ is the element of $\mathbb{R}^I$ such that $(u \circ v)_i = u_i v_i$. We denote by "$\cdot$" the (standard) scalar product in $\mathbb{R}^I : u \cdot v = \sum_{i \in I} u_i v_i$.

The function $\ln(\cdot) : \mathbb{R}^I_+ \to \mathbb{R}^I$ is the componentwise logarithm. That is, for each $u \in \mathbb{R}^I_+$, $(\ln u)_i = \ln(u_i)$. The function $\exp(\cdot) : \mathbb{R}^I \to \mathbb{R}^I_+$ is defined similarly: For each $u \in \mathbb{R}^I$, $(\exp u)_i = \exp(u_i)$.

If I is an arbitrary finite set, then by $\#(I)$ we mean the number of distinct elements in I.

2.2. Chemical reaction networks and their differential equations.

We begin with a definition of a chemical reaction network. By the *complexes* of a reaction network we shall mean the linear combinations of the species that appear before and after the reaction arrows. In network (1.1) there are six complexes: $2A, A_2, A_2 + B, A_2B, A_2B + C, A_2BC$. If $\mathscr{S}$ is the set of species of the network, we view the complexes of the network to be members of $\bar{\mathbb{R}}^{\mathscr{S}}_+$. We take a reaction network to be a specification of its species, a specification of its complexes, and, finally, a specification of a "reacts to" relation among the complexes.

DEFINITION 2.1. *A* chemical reaction network *consists of three finite sets:*

1. *A set $\mathscr{S}$ of distinct* species *of the network;*
2. *a set $\mathscr{C} \subset \bar{\mathbb{R}}_+^{\mathscr{S}}$ of* complexes *of the network;*
3. *a set $\mathscr{R} \subset \mathscr{C} \times \mathscr{C}$ of* reactions, *with the following properties:*
   (a) *$(y, y) \notin \mathscr{R}$ for any $y \in \mathscr{C}$;*
   (b) *for each $y \in \mathscr{C}$ there exists $y' \in \mathscr{C}$ such that $(y, y') \in \mathscr{R}$ or such that $(y', y) \in \mathscr{R}$.*

Following the usual notation in chemistry we write $y \to y'$ to indicate the reaction whereby complex y reacts to complex y'. With each reaction $y \to y'$ we associate the *reaction vector* $y' - y \in \mathbb{R}^{\mathscr{S}}$. In the context of our example, the reaction vector corresponding to $A_2B + C \to A_2BC$ is $A_2BC - A_2B - C$. For reasons that will become apparent, the span of a network's reaction vectors has special significance. This serves as motivation for the following definition.

DEFINITION 2.2. *The* stoichiometric subspace *of a reaction network $\{\mathscr{S}, \mathscr{C}, \mathscr{R}\}$ is the set $\mathrm{S} \subset \mathbb{R}^{\mathscr{S}}$ defined by*

$$\mathrm{S} = \operatorname{span}\left\{y' - y \in \mathbb{R}^{\mathscr{S}} : y \to y' \in \mathscr{R}\right\}.$$

When a particular network is under discussion, it will be understood that the symbol S is reserved to denote its stoichiometric subspace. We denote by $\mathrm{S}^\perp \subset \mathbb{R}^{\mathscr{S}}$ the orthogonal complement of S relative to the standard scalar product in $\mathbb{R}^{\mathscr{S}}$. We reserve the symbol p $(= \#(\mathscr{S}) - \dim \mathrm{S})$ for the dimension of $\mathrm{S}^\perp$.

Note that if $M \in \mathbb{R}_+^{\mathscr{S}}$ is the vector of molecular weights of the species in a network and if $y \to y'$ is a reaction, then $y \cdot M$ is the total mass of molecules on the reactant side of the reaction, while $y' \cdot M$ is the total mass of molecules on the product side. If the reaction respects conservation of mass, then we expect that $y \cdot M = y' \cdot M$, or equivalently, that $(y' - y) \cdot M = 0$. If all the reactions in the network respect conservation of mass, then M should be orthogonal to each of the reaction vectors, which is to say that M should be a member of $\mathrm{S}^\perp$. Following Horn and Jackson [16], we say that a reaction network is conservative if there exists for it at least one positive member of $\mathrm{S}^\perp$ that might play the role of a vector of molecular weights, relative to which all reactions are mass conserving.

DEFINITION 2.3. *A chemical reaction network $\{\mathscr{S}, \mathscr{C}, \mathscr{R}\}$ is* conservative *if $\mathrm{S}^\perp \cap \mathbb{R}_+^{\mathscr{S}} \neq \emptyset$.*

In order to write the differential equations governing the species concentrations in a reaction network, it is first necessary to specify, at each mixture composition, the rate at which the various reactions occur. For each reaction network $\{\mathscr{S}, \mathscr{C}, \mathscr{R}\}$ we generally denote by $c \in \bar{\mathbb{R}}_+^{\mathscr{S}}$ the vector of molar concentrations of the species. That is, for each $s \in \mathscr{S}$, c_s is the molar concentration of species s. By a *rate function* for the reaction $y \to y'$ we mean a function $\mathscr{K}_{y\to y'}(\cdot) : \bar{\mathbb{R}}_+^{\mathscr{S}} \to \bar{\mathbb{R}}_+$ that gives information about the dependence of occurrence rate on mixture composition. In particular, $\mathscr{K}_{y\to y'}(c)$ is the molar occurrence rate per unit volume at composition c. It is natural to expect that, at a particular mixture composition c, $\mathscr{K}_{y\to y'}(c)$ will be positive if and only if all species appearing with nonzero stoichiometric coefficients in the *reactant complex* y are present at composition c—that is, if and only if $\operatorname{supp}(y) \subset \operatorname{supp}(c)$. Very often, rate functions are taken to be of *mass-action* type—monomials in the species concentrations reflecting the probability of an encounter between molecules appearing in the reactant complex.

Definition 2.4. *A* kinetics $\mathscr{K}$ *for a reaction network* $\{\mathscr{S},\mathscr{C},\mathscr{R}\}$ *is an assignment to each reaction* $y \to y' \in \mathscr{R}$ *of a continuous rate function* $\mathscr{K}_{y\to y'}(\cdot) : \bar{\mathbb{R}}_+^{\mathscr{S}} \to \bar{\mathbb{R}}_+$ *such that* $\mathscr{K}_{y\to y'}(c) > 0$ *if and only if* $\operatorname{supp}(y) \subset \operatorname{supp}(c)$. *A* mass-action kinetics *for the network is a kinetics of the following kind: For each reaction* $y \to y' \in \mathscr{R}$ *there is a positive* rate constant $k_{y\to y'}$ *such that*

$$\mathscr{K}_{y\to y'}(c) \equiv k_{y\to y'} \prod_{s\in\mathscr{S}} (c_s)^{y_s}. \tag{2.1}$$

A kinetic system *is a reaction network endowed with a kinetics. A* mass-action system *is a reaction network endowed with a mass-action kinetics.*

Definition 2.5. *The* species-formation-rate function *for a kinetic system* $\{\mathscr{S}, \mathscr{C},\mathscr{R},\mathscr{K}\}$ *is the function* $r(\cdot) : \bar{\mathbb{R}}_+^{\mathscr{S}} \to \mathbb{R}^{\mathscr{S}}$ *defined by*

$$r(c) = \sum_{y\to y'\in\mathscr{R}} \mathscr{K}_{y\to y'}(c)(y' - y), \tag{2.2}$$

and the associated differential equation *is*

$$\dot{c} = r(c). \tag{2.3}$$

We say that $c \in \bar{\mathbb{R}}_+^{\mathscr{S}}$ *is an* equilibrium composition *if* $r(c) = 0$. *By the* positive equilibrium set *for the kinetic system we mean the set*

$$E = \left\{c \in \mathbb{R}_+^{\mathscr{S}} : r(c) = 0\right\}. \tag{2.4}$$

Example 2.6. Network (1.1), taken with mass action kinetics, gives rise to the species-formation-rate function (2.5), which is constructed according to (2.1) and (2.2):

$$\begin{aligned} r_A(c) &= 2k_{A_2\to 2A}(c_{A_2}) - 2k_{2A\to A_2}(c_A)^2,\\ r_B(c) &= k_{A_2B\to A_2+B}(c_{A_2B}) - k_{A_2+B\to A_2B}(c_{A_2})(c_B),\\ r_C(c) &= k_{A_2BC\to A_2B+C}(c_{A_2BC}) - k_{A_2B+C\to A_2BC}(c_{A_2B})(c_C),\\ r_{A_2}(c) &= k_{2A\to A_2}(c_A)^2 - k_{A_2\to 2A}(c_{A_2})\\ &\quad + k_{A_2B\to A_2+B}(c_{A_2B}) - k_{A_2+B\to A_2B}(c_{A_2})(c_B),\\ r_{A_2B}(c) &= k_{A_2+B\to A_2B}(c_{A_2})(c_B) - k_{A_2B\to A_2+B}(c_{A_2B})\\ &\quad + k_{A_2BC\to A_2B+C}(c_{A_2BC}) - k_{A_2B+C\to A_2BC}(c_{A_2B})(c_C),\\ r_{A_2BC}(c) &= k_{A_2B+C\to A_2BC}(c_{A_2B})(c_C) - k_{A_2BC\to A_2B+C}(c_{A_2BC}). \end{aligned} \tag{2.5}$$

Remark 2.7. Note that the species-formation-rate function takes values in the stoichiometric subspace S, which, for a conservative system, will be a *proper* linear subspace of $\mathbb{R}^{\mathscr{S}}$. Because in (2.3) the "velocity vector" $\dot{c}$ is restricted to point along S, it is not difficult to see that a composition trajectory that begins at composition c' can pass through composition c'' only if $c'' - c'$ lies in S.

With this in mind we say that two compositions are *stoichiometrically compatible* if $c'' - c'$ lies in S. Stoichiometric compatibility is an equivalence relation that serves to partition the set $\bar{\mathbb{R}}_+^{\mathscr{S}}$ of all compositions into *stoichiometric compatibility classes* (and partition the set $\mathbb{R}_+^{\mathscr{S}}$ of strictly positive compositions into *positive stoichiometric compatibility classes*). Stoichiometric compatibility classes are those subsets of parallels of S that lie in $\bar{\mathbb{R}}_+^{\mathscr{S}}$. In particular, the stoichiometric compatibility class containing a composition c^0 is the intersection of $\bar{\mathbb{R}}_+^{\mathscr{S}}$ with the parallel $c^0 + \mathrm{S} = \{c^0 + \sigma \in \mathbb{R}^{\mathscr{S}} : \sigma \in \mathrm{S}\}$.

Because two compositions along a trajectory governed by (2.3) must be stoichiometrically compatible, it is evident that each composition trajectory must lie entirely within a stoichiometric compatibility class.

2.3. Quasi-thermostatic kinetic systems.

It is remarkable that a wide variety of mass-action systems fall into a class called by Horn and Jackson [16] *quasi-thermostatic*. Quasi-thermostatic systems are characterized by a simply described positive equilibrium set.

DEFINITION 2.8. *A kinetic system* $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{K}\}$ *is called* quasi-thermostatic *if it admits a positive equilibrium* $c^* \in \mathbb{R}_+^{\mathscr{S}}$ *and if its positive equilibrium set is given by*

$$E = \left\{c \in \mathbb{R}_+^{\mathscr{S}} : \ln c - \ln c^* \in \mathrm{S}^{\perp}\right\}. \tag{2.6}$$

Remark 2.9. Given the stoichiometric subspace for a quasi-thermostatic system, specification of its entire positive equilibrium set amounts to the specification of a single member, c^*. In fact, if c^{**} is any other member, then the positive equilibrium set has the alternative characterization

$$E = \left\{c \in \mathbb{R}_+^{\mathscr{S}} : \ln c - \ln c^{**} \in \mathrm{S}^{\perp}\right\}.$$

Remark 2.10. Not all mass-action systems are quasi thermostatic [16, 9, 2, 3, 4], but very important categories of them are. In particular, *detailed balanced* mass-action systems are quasi thermostatic [16]. In the terminology of chemistry, a kinetic system $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{K}\}$ is detailed balanced if all of its reactions are reversible (i.e., $y \to y'$ implies $y' \to y$) and if for each $y \to y' \in \mathscr{R}$ and at each equilibrium $c^* \in \mathbb{R}_+^{\mathscr{S}}$,

$$\mathscr{K}_{y \to y'}(c^*) = \mathscr{K}_{y' \to y}(c^*).$$

Not all reversible mass-action systems are detailed balanced: In some instances, reaction network structure alone will suffice to ensure detailed balancing, regardless of rate-constant values, but in other instances detailed balancing obtains only if the rate constants are suitably orchestrated [10]. It is a commonly held belief among chemists that closed mass-action systems occurring in nature *should* be detailed balanced (and, therefore, quasi thermostatic). This belief seems to go back to Wegscheider [26] and Lewis [20].

Horn and Jackson [16] showed that mass-action systems satisfying the far weaker condition of *complex balancing* are also quasi thermostatic. (Detailed balancing implies complex balancing, but the converse is not true.) In turn, Feinberg [5] and Horn [17] showed that for a large class of mass-action systems (those that derive from *weakly reversible deficiency zero* networks), complex balancing, and therefore quasi-thermostatic behavior, is a consequence of reaction network structure alone, independent of rate-constant values. Subsequently, Feinberg [7, 8, 11] showed that, regardless of rate-constant values, quasi-thermostatic behavior (but not necessarily complex balancing) is a property of mass-action systems that derive from a still wider class of networks, those satisfying the requirements of the *deficiency one theorem.*

The following proposition tells us that, for a quasi-thermostatic kinetic system, each positive stoichiometric compatibility class (Remark 2.7) contains exactly one member of the positive equilibrium set. A proof can be found in [16], and a somewhat different proof is given in Appendix B of [11].

PROPOSITION 2.11. *Let $\{\mathscr{S},\mathscr{C},\mathscr{R}\}$ be a reaction network with stoichiometric subspace* S*, and let c^*, c^0 be elements of $\mathbb{R}^{\mathscr{S}}_+$. Then the sets*

$$c^0 + \mathrm{S} = \left\{c^0 + \sigma \in \mathbb{R}^{\mathscr{S}} : \sigma \in \mathrm{S}\right\}$$

and

$$E = \left\{c \in \mathbb{R}^{\mathscr{S}}_+ : \ln c - \ln c^* \in \mathrm{S}^{\perp}\right\}$$

meet at exactly one point.

Remark 2.12. If, for a quasi-thermostatic system, we write the equilibrium set in the form

$$E = \left\{c \in \mathbb{R}^{\mathscr{S}}_+ : c = c^* \circ \exp(\gamma),\ \gamma \in \mathrm{S}^{\perp}\right\}, \tag{2.7}$$

it becomes clear that E is a p-dimensional smooth manifold ($p = \dim \mathrm{S}^{\perp}$) parametrized globally by the map $\tilde{c}(\cdot) : \mathrm{S}^{\perp} \to \mathbb{R}^{\mathscr{S}}$, where $\tilde{c}(\gamma) \equiv c^* \circ \exp(\gamma)$. (After choosing a basis for $\mathrm{S}^{\perp}$, one can of course identify $\mathrm{S}^{\perp}$ with $\mathbb{R}^p$.) Clearly, $\tilde{c}(0) = c^*$. For future reference we note that the derivative of $\tilde{c}(\cdot)$, evaluated at 0, and denoted $\mathrm{d}\,\tilde{c}[0] : \mathrm{S}^{\perp} \to \mathbb{R}^{\mathscr{S}}$, acts on each $\gamma \in \mathrm{S}^{\perp}$ in the following way: $\mathrm{d}\,\tilde{c}[0]\gamma = c^* \circ \gamma$. From this it is apparent that the linear subspace tangent to E at c^* is

$$c^* \circ \mathrm{S}^{\perp} = \left\{c^* \circ \gamma : \gamma \in \mathrm{S}^{\perp}\right\}. \tag{2.8}$$

3. Elemented reaction networks. By an *elemented reaction network* we mean a reaction network taken together with a distinguished choice of nonnegative basis for $\mathrm{S}^{\perp}$. As will be readily apparent, each basis vector gives rise to a linear combination of the species concentrations that is conserved along solutions of the differential equations associated with any kinetic system deriving from the network. Loosely speaking, then, each basis vector can be associated with an "indestructible element" whose total concentration is conserved by the reactions. In some instances, as in the case of network (1.1), these conserved elements (A, B, and C in our example) have natural interpretations and correspond to a natural choice of basis for $\mathrm{S}^{\perp}$. For a different network, the choice of elements might be more arbitrary and may, in fact, vary from one application to another.

As we shall see later, whatever the choice of elements, specification of their total (time-invariant) concentrations will serve to specify, bijectively, a particular stoichiometric compatibility class (Remark 2.7). For a quasi-thermostatic kinetic system, each positive stoichiometric compatibility class contains precisely one equilibrium (Proposition 2.11). In effect, then, the positive equilibrium set E can be parametrized by the total concentrations of the elements. As suggested in the introduction, our interest is in how the equilibria change in response to a change in the element concentrations, which, in fact, corresponds to a change in stoichiometric compatibility class.

DEFINITION 3.1. *An* elemented reaction network *consists of the following:*

1. *A reaction network $\{\mathscr{S},\mathscr{C},\mathscr{R}\}$;*
2. *a set $\mathscr{E}$ of p $(= \dim \mathrm{S}^{\perp})$ distinct members called the* elements *of the network;*
3. *a basis $\mathscr{M} = \{M_e\}_{e \in \mathscr{E}} \subset \bar{\mathbb{R}}^{\mathscr{S}}_+$ for $\mathrm{S}^{\perp}$.*

For each $e \in \mathscr{E}$ and $s \in \mathscr{S}$, the component M^s_e is called the e-content of species s*.*

Remark 3.2. Although we have not built such a requirement into the definition, we shall, in this article, suppose that all networks under consideration are conservative in the sense of Definition 2.3. (There are instances, not considered here, in which one might want to study the more general situation.) It is easy to see that, for a conservative network, it is always possible to choose a basis for $\mathrm{S}^{\perp}$ that lies in $\bar{\mathbb{R}}_{+}^{\mathscr{S}}$. Moreover, it is not difficult to see that, for an elemented conservative network, there exists for each species s at least one element e such that the e-content of s is not zero.

Example 3.3. One elemented network derived from (1.1) results from the choice $\mathscr{E} = \{A, B, C\}$ and the basis for $\mathrm{S}^{\perp}$ given by

$$\begin{aligned} M_A &= A + 2A_2 + 2A_2B + 2A_2BC, \\ M_B &= B + A_2B + A_2BC, \\ M_C &= C + A_2BC. \end{aligned} \tag{3.1}$$

In this case all of the elements are associated with certain species that appear explicitly in the network.

Example 3.4. It is instructive to consider a simple network of the general form

$$W + X \rightleftharpoons Y + Z, \tag{3.2}$$

for which $\dim \mathrm{S}^{\perp} = 3$. The network might be "elemented" in different ways, depending on the context. If, for example, $W = AQ_2$, $Y = AQ$, and $Z = XQ$, then (3.2) represents a transfer of a molecular component Q from species AQ_2 to species X. In this case, it is natural to take $\mathscr{E} = \{A, Q, X\}$ and

$$\begin{aligned} M_A &= AQ_2 + AQ = W + Y, \\ M_Q &= 2AQ_2 + AQ + XQ = 2W + Y + Z, \\ M_X &= X + XQ = X + Z. \end{aligned}$$

Note that the element X is an overt species, whereas the elements A and Q are not.

If, on the other hand, $W = AB$, $X = CD$, $Y = AC$, and $Z = BD$, then (3.2) represents a component exchange. One natural choice of elements is $\mathscr{E} = \{A, B, C\}$ and the basis for $\mathrm{S}^{\perp}$ given by

$$\begin{aligned} M_A &= AB + AC = W + Y, \\ M_B &= AB + BD = W + Z, \\ M_C &= CD + AC = X + Y. \end{aligned}$$

Still another choice is $\mathscr{E} = \{A, B, D\}$ and

$$\begin{aligned} M_A &= AB + AC = W + Y, \\ M_B &= AB + BD = W + Z, \\ M_D &= CD + BD = X + Z. \end{aligned}$$

With neither choice does any element appear explicitly as a species.

Recall that for the network discussed in the introduction we considered the *total concentrations* of the three elements. For each element e, the total concentration was calculated from the current mixture composition c by a sum of the following kind:

$$\sum_{s \in \mathscr{S}} M_e^s c_s,$$

where M_e^s is the e-content of species s. We now extend this idea using the following definition.

DEFINITION 3.5. *Let* $\{\mathscr{S},\mathscr{C},\mathscr{R},\mathscr{E},\mathscr{M}\}$ *be an elemented reaction network. The* element concentration functions $\bar{T}_e(\cdot):\mathbb{R}^{\mathscr{S}}\to\mathbb{R}$ *are defined for each* $e\in\mathscr{E}$ *by*

$$\bar{T}_e(x) := M_e \cdot x. \tag{3.3}$$

For use later, we define the function $\bar{T}(\cdot):\mathbb{R}^{\mathscr{S}}\to\mathbb{R}^{\mathscr{E}}$ by

$$\bar{T}(x) := \sum_{e\in\mathscr{E}} \bar{T}_e(x)e. \tag{3.4}$$

It is not difficult to see that the kernel of $\bar{T}(\cdot)$ is the stoichiometric subspace S.

Remark 3.6. Two compositions c' and c'' are associated with precisely the same element concentrations (i.e., $\bar{T}(c')=\bar{T}(c'')$) if and only if $c'-c''$ is a member of the stoichiometric subspace, which is to say that c' and c'' reside in the same stoichiometric compatibility class. In this way each specification of positive element concentrations (i.e., each member of $\bar{T}(\mathbb{R}_+^{\mathscr{S}})\subset\mathbb{R}_+^{\mathscr{E}}$) can be identified bijectively with a positive stoichiometric compatibility class.

DEFINITION 3.7. *An* elemented kinetic system *is an elemented reaction network endowed with kinetics.*

Remark 3.8. It is apparent from Remark 3.6 that the element concentrations remain invariant along solutions of the differential equations associated with an elemented kinetic system: Composition trajectories reside entirely within stoichiometric compatibility classes (Remark 2.7), and, within each stoichiometric compatibility class, all compositions give rise to the same element concentrations. The constancy of the element concentrations can, of course, be seen more directly: If $c(\cdot)$ is a solution of the differential equation for a particular elemented kinetic system, and if e is an element, then, at each t,

$$\frac{d}{dt}\bar{T}_e(c(t)) = M_e\cdot\frac{dc(t)}{dt} = M_e\cdot r(c(t)) = 0. \tag{3.5}$$

Here, of course, $M_e\in \mathrm{S}^{\perp}$ is the basis vector associated with element e, and $r(\cdot)$ is the species-formation-rate function for the kinetic system. The last equality in (3.5) follows from the fact that $r(\cdot)$ takes values in S.

4. Elemented quasi-thermostatic systems. We asserted in Remark 3.6 that, for an elemented network, there is a bijective correspondence between positive stoichiometric compatibility classes and distinct positive specifications of element concentrations. On the other hand, for a quasi-thermostatic system, each positive stoichiometric compatibility class contains precisely one equilibrium (Proposition 2.11). Thus, for a quasi-thermostatic elemented system, there is a bijective correspondence between the positive equilibrium set E and $\bar{T}(\mathbb{R}_+^{\mathscr{S}})$, the set of all (realizable) positive specifications of the element concentrations. Our goal in this section is to show that this correspondence is, in fact, a diffeomorphism. Then we will be able to speak of the smooth dependence of equilibrium species concentrations on the element concentrations, and the way will be paved for discussion and analysis of sensitivities, which were described only informally in the introduction.

We consider, therefore, a quasi-thermostatic elemented kinetic system $\{\mathscr{S},\mathscr{C},\mathscr{R},\mathscr{K},\mathscr{E},\mathscr{M}\}$ with positive equilibrium set $E\in\mathbb{R}_+^{\mathscr{S}}$. We denote by $\bar{T}_E(\cdot):E\to\bar{T}(\mathbb{R}_+^{\mathscr{S}})$ the restriction of $\bar{T}(\cdot)$ to E, taken with the indicated choice of codomain. (It is the function $\bar{T}_E(\cdot)$ that we want to show is a diffeomorphism.) Note that $\bar{T}_E(\cdot)$ is a map from the p-dimensional smooth manifold E to the p-dimensional smooth manifold $\bar{T}(\mathbb{R}_+^{\mathscr{S}})$ (an open set in $\mathbb{R}^{\mathscr{E}}$ by the open mapping theorem [1]). Recall from Remark 2.12 that if c^* is a member of E, then at c^* the space tangent to E is

$$c^* \circ \mathrm{S}^{\perp} = \left\{ c^* \circ \gamma : \gamma \in \mathrm{S}^{\perp} \right\}.$$

NOTATION 4.1. *Occasionally we will focus on a fixed but arbitrary positive equilibrium $c^* \in E$. It will be useful to have available a scalar product "$*$" in $\mathbb{R}^{\mathscr{S}}$ defined by*

$$x * y = x \cdot (c^* \circ y) \equiv \sum_{s \in \mathscr{S}} x_s c_s^* y_s.$$

LEMMA 4.2. *The derivative of $\bar{T}_E(\cdot)$ at an equilibrium $c^* \in E$ is the nonsingular map $\mathrm{d}\,\bar{T}_E\,[c^*]\,(\cdot) : c^* \circ \mathrm{S}^{\perp} \to \mathbb{R}^{\mathscr{E}}$ given for each $\gamma \in \mathrm{S}^{\perp}$ by*

$$\mathrm{d}\,\bar{T}_E\,[c^*]\,(c^* \circ \gamma) = \sum_{e \in \mathscr{E}} (M_e * \gamma) e. \tag{4.1}$$

Proof. Equation (4.1) derives from straightforward calculation. To see that the map $\mathrm{d}\,\bar{T}_E\,[c^*]\,(\cdot)$ is nonsingular, suppose that $\gamma^{\#} \in \mathrm{S}^{\perp}$ is such that $\mathrm{d}\,\bar{T}_E\,[c^*]\,(c^* \circ \gamma^{\#}) = 0$. This requires that $M_e * \gamma^{\#} = 0$ for each $e \in \mathscr{E}$. Thus, $\gamma^{\#} \in \mathrm{S}^{\perp}$ is orthogonal, with respect to the "$*$" scalar product, to each member of a basis for $\mathrm{S}^{\perp}$. Hence, $\gamma^{\#}$ and therefore $c^* \circ \gamma^{\#}$ are each the zero vector in $\mathbb{R}^{\mathscr{S}}$. □

PROPOSITION 4.3. *$\bar{T}_E(\cdot) : E \to \bar{T}(\mathbb{R}_+^{\mathscr{S}})$ is a diffeomorphism.*

Proof. To show that $\bar{T}_E(\cdot)$ is bijective, suppose that T^0 is some fixed but arbitrary member of $\bar{T}(\mathbb{R}_+^{\mathscr{S}})$. We will argue that there is precisely one member of E mapped to the point T^0 by $\bar{T}_E(\cdot)$. Because T^0 is a member of $\bar{T}(\mathbb{R}_+^{\mathscr{S}})$, there is a $c^0 \in \mathbb{R}_+^{\mathscr{S}}$ such that $\bar{T}(c^0) = T^0$. That is,

$$\bar{T}(c^0) = \sum_{e \in \mathscr{E}} (M_e \cdot c^0) e = T^0.$$

In fact, the full set of vectors in $\mathbb{R}^{\mathscr{S}}$ mapped by $\bar{T}(\cdot)$ to T^0 is $c^0 + \ker \bar{T}(\cdot) = c^0 + \mathrm{S}$. Our interest, then, is in the members of $c^0 + \mathrm{S}$ that are also members of E. If c^* is a member of E, then, for the quasi-thermostatic system under study, the equilibrium set is given by

$$E = \left\{ c \in \mathbb{R}_+^{\mathscr{S}} : \ln c - \ln c^* \in \mathrm{S}^{\perp} \right\}.$$

From Proposition 2.11, however, E meets $c^0 + \mathrm{S}$ in *exactly* one point. Thus, $\bar{T}_E(\cdot)$ is bijective.

Because from Lemma 4.2 we have that $\mathrm{d}\,\bar{T}_E\,[c^*]\,(\cdot)$ is nonsingular for each $c^* \in E$, the inverse function theorem [12] ensures that, at each point in $\bar{T}(\mathbb{R}_+^{\mathscr{S}})$, there is a local inverse of the smooth function $\bar{T}_E(\cdot)$ that is also smooth. Since any such local inverse must be a restriction of the (unique) global inverse whose existence was established above, it follows that the inverse of $\bar{T}(\cdot)$ is smooth. □

5. Sensitivities. Hereafter we consider an elemented quasi-thermostatic kinetic system $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{K}, \mathscr{E}, \mathscr{M}\}$, and we denote by c^* a fixed but arbitrary member of its positive equilibrium set E. We denote by $T^* \in \mathbb{R}^{\mathscr{E}}$ the value $\bar{T}(c^*)$. That is, T^* is the vector of element concentrations associated with the equilibrium composition c^*.

Proposition 4.3 tells us that the map $\bar{T}_E(\cdot) : E \to \bar{T}(\mathbb{R}^{\mathscr{S}}_+)$ gives a global coordinate system on E. That is, each equilibrium in E is characterized uniquely by a specification of the concentrations of the elements. We will choose to introduce a change in coordinates so that we can work instead with logarithms of the element concentrations. For this purpose we introduce the map $\ln \bar{T}_E(\cdot) : E \to \ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+)$ defined in the obvious way: For each $c \in E$, $\ln \bar{T}_E(c) = \ln(\bar{T}_E(c))$. As a variant of (4.1), it is not difficult to see that $\mathrm{d} \ln \bar{T}_E[c^*](\cdot) : c^* \circ \mathrm{S}^\perp \to \mathbb{R}^{\mathscr{E}}$ is given by

$$\mathrm{d} \ln \bar{T}_E[c^*](c^* \circ \gamma) = \sum_{e \in \mathscr{E}} \frac{1}{T^*_e}(M_e \cdot c^* \circ \gamma)e = \sum_{e \in \mathscr{E}} \frac{1}{T^*_e}(M_e * \gamma)e \tag{5.1}$$

for all $\gamma \in \mathrm{S}^\perp$.

Because $\ln \bar{T}_E(\cdot)$ amounts to the composition of the diffeomorphisms $\bar{T}_E(\cdot)$ and $\ln(\cdot) : \bar{T}(\mathbb{R}^{\mathscr{S}}_+) \to \ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+)$, it too is a diffeomorphism. Hereafter, we denote by $\bar{c}(\cdot) : \ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+) \to E$ the inverse of $\ln \bar{T}_E(\cdot)$. Thus, if $\ln T$ is a member of $\ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+)$ it makes sense to speak of "the equilibrium composition $\bar{c}(\ln T)$." Note that $\ln \bar{T}_E(c^*) = \ln \bar{T}(c^*) = \ln T^*$ and $\bar{c}(\ln T^*) = c^*$. Note too that $\mathrm{d}\,\bar{c}[\ln T^*](\cdot) : \mathbb{R}^{\mathscr{E}} \to c^* \circ \mathrm{S}^\perp$ is just the inverse of $\mathrm{d} \ln \bar{T}_E[c^*]$, which is given by (5.1). Thus we have

$$\mathrm{d}\,\bar{c}[\ln T^*]\left(\sum_{e \in \mathscr{E}} \frac{1}{T^*_e}(M_e * \gamma)e\right) = c^* \circ \gamma \tag{5.2}$$

for all $\gamma \in \mathrm{S}^\perp$.

We will also be interested in the map $\ln \bar{c}(\cdot) : \ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+) \to \ln E$ $(= \ln c^* + \mathrm{S}^\perp)$ defined in the obvious way. (Note that the tangent spaces of the manifolds $\ln \bar{T}(\mathbb{R}^{\mathscr{S}}_+)$ and $\ln c^* + \mathrm{S}^\perp$ are, respectively, everywhere $\mathbb{R}^{\mathscr{E}}$ and $\mathrm{S}^\perp$.) Because $\mathrm{d} \ln \bar{c}[\ln T^*](\cdot) : \mathbb{R}^{\mathscr{E}} \to \mathrm{S}^\perp$ is given by $\frac{1}{c^*} \circ \mathrm{d}\,\bar{c}[\ln T^*]$, it follows from (5.2) that

$$\mathrm{d} \ln \bar{c}[\ln T^*]\left(\sum_{e \in \mathscr{E}} \frac{1}{T^*_e}(M_e * \gamma)e\right) = \gamma \tag{5.3}$$

for all $\gamma \in \mathrm{S}^\perp$.

We are now in position to make precise definitions of the sensitivities discussed informally in the introduction.

Definition 5.1. *By the* sensitivity vector for element e at equilibrium composition c^* *we mean the vector in* $c^* \circ \mathrm{S}^\perp \subset \mathbb{R}^{\mathscr{S}}$ *given by*

$$\frac{\partial \ln \bar{c}}{\partial \ln T_e}(c^*) := \mathrm{d} \ln \bar{c}\,[\ln T^*]\,e. \tag{5.4}$$

By the sensitivity matrix at c^* *we mean the array whose elements are given by*

$$\left(\frac{\partial \ln \bar{c}_s}{\partial \ln T_e}(c^*)\right)_{s \in \mathscr{S},\, e \in \mathscr{E}}. \tag{5.5}$$

6. Means to calculate the sensitivity matrix. Here we provide means to calculate the sensitivity matrix at the equilibrium $c^* \in E$.[1] For the specified equilibrium c^*, we denote by $\{M^e\}_{e \in \mathscr{E}} \subset \mathrm{S}^\perp$ the basis for $\mathrm{S}^\perp$ that is reciprocal, relative

[1] We note that these calculations have points of contact with issues raised some years ago in two Ph.D. theses, by K. Israel [18] and J. Nailor [21], supervised by Michael Reed. Stated in very rough and somewhat inaccurate terms, their interest was in signs of entries in what we call the sensitivity matrix—signs that might be determined by experiment. The motivation was that such experimentally determined signs might provide clues about the underlying reaction network when the network is unknown.

to the "$*$" scalar product in $\mathbb{R}^{\mathscr{S}}$, to the basis $\mathscr{M} = \{M_e\}_{e\in\mathscr{E}}$. (Recall Notation 4.1.) That is, $M^{e'} * M_e = 1$ if $e' = e$ and is 0 otherwise. Because the "$*$" scalar product depends upon c^*, so will the reciprocal basis it induces. It is not difficult to determine such reciprocal bases computationally within the context of readily available symbolic mathematics programs.

THEOREM 6.1. *Let $\{\mathscr{S},\mathscr{C},\mathscr{R},\mathscr{K},\mathscr{E},\mathscr{M}\}$ be an elemented quasi-thermostatic kinetic system, let $c^* \in E$ be a positive equilibrium (corresponding to an element specification $T^* \in \mathbb{R}^{\mathscr{E}}$), and let $\{M^e\}_{e\in\mathscr{E}}$ be the basis for $\mathrm{S}^\perp$ reciprocal to $\mathscr{M}$ relative to the "$*$" scalar product in $\mathbb{R}^{\mathscr{S}}$. Then for each $e \in \mathscr{E}$, the sensitivity vector is given by*

$$\frac{\partial \ln \bar{c}}{\partial \ln T_e}(c^*) = T_e^* M^e. \tag{6.1}$$

Entries of the sensitivity matrix are given by

$$\frac{\partial \ln \bar{c}_s}{\partial \ln T_e}(c^*) = T_e^* M^{es} \tag{6.2}$$

for each $s \in \mathscr{S}$ and $e \in \mathscr{E}$.

Proof. Equation (6.1) results from a simple substitution in (5.3): To obtain the sensitivity vector for a particular element $e' \in \mathscr{E}$, set $\gamma = T_{e'}^* M^{e'}$. Equation (6.2) is just the component form of (6.1). □

Example 6.2. Consider network (6.3), which can serve to model a process whereby an active biomolecule A_2B_2 is assembled from elements A and B:

$$A + B \rightleftharpoons AB, \qquad 2AB \rightleftharpoons A_2B_2. \tag{6.3}$$

One elemented network derived from (6.3) is afforded by the choices $\mathscr{E} = \{A, B\}$ and $\mathscr{M} = \{M_A, M_B\}$, where

$$\begin{aligned} M_A &= A + AB + 2A_2B_2,\\ M_B &= B + AB + 2A_2B_2. \end{aligned} \tag{6.4}$$

If the kinetics for the network is mass action, then, for any choice of rate constants, detailed balance will obtain [10] and the system will be quasi thermostatic. Thus, Theorem 6.1 can be used to calculate all entries in the sensitivity matrix, evaluated at an arbitrary equilibrium c^*. In particular, the entries corresponding to the "active" species A_2B_2 are

$$\begin{aligned} \frac{\partial \ln \bar{c}_{A_2B_2}}{\partial \ln T_A}(c^*) &= \frac{2(c_A^* + c_{AB}^* + 2c_{A_2B_2}^*)c_B^*}{c_A^*c_B^* + (c_A^* + c_B^*)(c_{AB}^* + 4c_{A_2B_2}^*)},\\ \frac{\partial \ln \bar{c}_{A_2B_2}}{\partial \ln T_B}(c^*) &= \frac{2(c_B^* + c_{AB}^* + 2c_{A_2B_2}^*)c_A^*}{c_A^*c_B^* + (c_A^* + c_B^*)(c_{AB}^* + 4c_{A_2B_2}^*)}. \end{aligned} \tag{6.5}$$

7. Degree, connectivity, and network-imposed sensitivity bounds. As before, we consider an elemented quasi-thermostatic kinetic system $\{\mathscr{S},\mathscr{C},\mathscr{R},\mathscr{K},\mathscr{E},\mathscr{M}\}$, and we let c^* be a positive equilibrium, corresponding to an element concentration specification $T^* \in \mathbb{R}^{\mathscr{E}}$. Motivated by considerations discussed in the introduction, we have interest in understanding the extent to which the equilibrium concentration of a particular species $s \in \mathscr{S}$ might be robust against variations in the supply of elements. Our interest, then, is in the "worst case"—that is, in the most sensitive response in the concentration of s, taken against variation in each of the element concentrations.

Definition 7.1. *The* sensitivity of species $s \in \mathscr{S}$ at c^*, *denoted* $\Lambda^s(c^*)$, *is defined by*

$$\Lambda^s(c^*) = \max_{e \in \mathscr{E}} \left\{ \left| \frac{\partial \ln \bar{c}_s}{\partial \ln T_e}(c^*) \right| \right\}. \tag{7.1}$$

By the system sensitivity at c^*, *denoted* $\Lambda(c^*)$, *we mean the least of the species sensitivities. That is,*

$$\Lambda(c^*) = \min_{s \in \mathscr{S}} \Lambda^s(c^*). \tag{7.2}$$

Remark 7.2. In general, the species sensitivities and the system sensitivity will depend on the choice of basis $\mathscr{M}$ for $\mathrm{S}^{\perp}$.

If the system sensitivity is high at c^*, then *all* species sensitivities at c^* are high, so it makes sense to say that the system itself is highly sensitive to element concentration variations.

Our goal in this section is to provide lower bounds on the species sensitivities and on the system sensitivity, bounds that derive *from network properties alone*, independent of kinetic parameters and even of the equilibrium at which the sensitivities are calculated. By way of preparation, we posit formally some ideas that were discussed informally in the introduction. All the following are attributes of the elemented *network* underlying the kinetic system we have been studying:

For each $e \in \mathscr{E}$ we let

$$M_e^{max} = \max_{s \in \mathscr{S}} \{M_e^s\}. \tag{7.3}$$

That is, M_e^{max} is the largest number of copies of element e that can be found in any species. It is a measure of the extent to which element e combines in a multiplicitous way.

For each $e, e' \in \mathscr{E}$ we let

$$M_e^{max}(e') = \max_{s \in \operatorname{supp} M_{e'}} \{M_e^s\}. \tag{7.4}$$

That is, $M_e^{max}(e')$ is the largest number of copies of element e that can be found in any species *that also contains* e'. Clearly we have

$$M_e^{max}(e') \leq M_e^{max} \tag{7.5}$$

for each $e, e' \in \mathscr{E}$.

By the *degree* of element e we mean

$$\deg(e) = \sum_{e' \in \mathscr{E}} M_e^{max}(e'). \tag{7.6}$$

As we indicated in the introduction, an element will have high degree if it combines gregariously with many different elements or if it combines in high multiplicity with at least one of the elements.

THEOREM 7.3. *Let $\{\mathscr{S},\mathscr{C},\mathscr{R},\mathscr{K},\mathscr{E},\mathscr{M}\}$ be an elemented quasi-thermostatic kinetic system with positive equilibrium set E. For each $c^* \in E$ and each $s \in \mathscr{S}$,*

$$\Lambda^s(c^*) \geq \max_{e\in\mathscr{E}} \left\{ \frac{M_e^s}{\deg(e)} \right\}. \tag{7.7}$$

Moreover,

$$\Lambda(c^*) \geq \min_{s\in\mathscr{S}} \max_{e\in\mathscr{E}} \left\{ \frac{M_e^s}{\deg(e)} \right\}. \tag{7.8}$$

Proof. We begin with the equation

$$M_e = \mathrm{d}\ln\bar{c}\,[\ln T^*] \left(\sum_{e'\in\mathscr{E}} \frac{1}{T^*_{e'}} (M_{e'} * M_e) e' \right), \tag{7.9}$$

which is an immediate consequence of (5.3). Written in terms of components, (7.9) gives for each $e \in \mathscr{E}$ and $s \in \mathscr{S}$,

$$M_e^s = \sum_{e'\in\mathscr{E}} \frac{\partial \ln \bar{c}_s}{\partial \ln T_{e'}}(c^*) \frac{1}{T^*_{e'}} (M_{e'} * M_e). \tag{7.10}$$

From this we can write the inequality

$$M_e^s \leq \sum_{e'\in\mathscr{E}} \left| \frac{\partial \ln \bar{c}_s}{\partial \ln T_{e'}}(c^*) \right| \frac{1}{T^*_{e'}} (M_{e'} * M_e). \tag{7.11}$$

Note that for each $e', e \in \mathscr{E}$ we have

$$\begin{aligned} M_{e'} * M_e = \sum_{s\in\mathscr{S}} M^s_{e'} c^*_s M^s_e &= \sum_{s\in \operatorname{supp} M_{e'}} M^s_{e'} c^*_s M^s_e \\ &\leq M_e^{max}(e') \sum_{s\in \operatorname{supp} M_{e'}} M^s_{e'} c^*_s = M_e^{max}(e') T^*_{e'}. \end{aligned} \tag{7.12}$$

Thus, from inequalities (7.11) and (7.12) and equations (7.1) and (7.6) we obtain for every $e \in \mathscr{E}$ and $s \in \mathscr{S}$ the inequality

$$M_e^s \leq \Lambda^s(c^*) \sum_{e'\in\mathscr{E}} M_e^{max}(e') = \Lambda^s(c^*) \deg(e), \tag{7.13}$$

or equivalently,

$$\Lambda^s(c^*) \geq \frac{M_e^s}{\deg(e)}. \tag{7.14}$$

Since (7.14) holds for every element $e \in \mathscr{E}$, we in fact have, for each $s \in \mathscr{S}$, the bound given in (7.7). The bound (7.8) is then just a consequence of the definition of system sensitivity. □

Example 7.4. In the case of the elemented kinetic system of Example 6.2, we have from (6.4), (7.4), and (7.6) that

$$\deg(A) = 4, \qquad \deg(B) = 4. \tag{7.15}$$

From (6.4), (7.7), and (7.15) we obtain

$$\begin{aligned} \Lambda^{A}(c^*) &\geq \frac{1}{4}, & \Lambda^{B}(c^*) &\geq \frac{1}{4}, \\ \Lambda^{AB}(c^*) &\geq \frac{1}{4}, & \Lambda^{A_2B_2}(c^*) &\geq \frac{1}{2}. \end{aligned} \tag{7.16}$$

Note that the lower bound on the "active" species A_2B_2 is sharp: If, for example, $c^* = \epsilon A + \epsilon B + \epsilon AB + A_2B_2$, then, from (6.5) and (7.1), we have that $\Lambda^{A_2B_2}(c^*) = \frac{1}{2}(1+\epsilon) + o(\epsilon^2)$. Thus, as ϵ tends to zero, $\Lambda^{A_2B_2}$ tends to the lower bound $\frac{1}{2}$ in (7.16). We note that for any choice of c^* there is an assignment of rate constants for network (6.2) consistent with the existence of the equilibrium c^* and at which detailed balance obtains. Thus, there are quasi-thermostatic systems for which, at certain equilibria, the sensitivity of A_2B_2 approaches the stipulated lower bound arbitrarily closely.

Remark 7.5. In a conservative elemented network, for each species $s \in \mathscr{S}$ there exists an element $e \in \mathscr{E}$ such that the e-content of s is positive (Remark 3.2). Thus, inequality (7.7) implies that a species (or system) sensitivity of zero is impossible in conservative quasi-thermostatic systems. It should be noted, however, that kinetic systems exist that *do* allow for zero sensitivity. One such system, proposed in the context of bacterial signaling, is treated in [24]. In this system the quasi-thermostatic condition is *not* satisfied.

Remark 7.6. Because the sensitivities themselves will depend on the choice of elements, so will the bounds given by Theorem 7.3. Consider, for example, the very simple network

$$2A + B \rightleftharpoons A_2B. \tag{7.17}$$

This network can be elemented in various ways: One choice is given by

$$\begin{aligned} M_A &= A + 2A_2B, \\ M_B &= B + A_2B; \end{aligned} \tag{7.18}$$

another equally valid choice is given by

$$\begin{aligned} M_Q &= A + \epsilon B + (2+\epsilon)A_2B, \\ M_R &= A + 2\epsilon B + (2+2\epsilon)A_2B, \end{aligned} \tag{7.19}$$

where ϵ is a positive number much smaller than 1. From Theorem 7.3 we have that, for choice (7.18), the species sensitivities are bounded from below according to the inequalities $\Lambda^A(c^*) \geq \frac{1}{4}$, $\Lambda^B(c^*) \geq \frac{1}{2}$, and $\Lambda^{A_2B}(c^*) \geq \frac{1}{2}$. For the system sensitivity we have that $\Lambda(c^*) \geq \frac{1}{4}$. On the other hand, for choice (7.19) the same theorem gives the bounds $\Lambda^A(c^*) \geq \frac{1}{4} + o(\epsilon)$, $\Lambda^B(c^*) \geq o(\epsilon)$, $\Lambda^{A_2B}(c^*) \geq \frac{1}{2}$, and $\Lambda(c^*) \geq o(\epsilon)$. Thus, different choices of basis for $\mathrm{S}^{\perp}$ can lead to markedly different lower bounds. Note that while choice (7.18) carries a natural interpretation for the elements A and B as building blocks that appear either overtly as species or latently in the compound A_2B, choice (7.19) offers no immediate physical interpretation for the elements Q and R.

Note also, that in the general case, if each vector in the basis for $\mathrm{S}^{\perp}$ is multiplied by a (potentially different) scalar, then the lower bounds corresponding to the new basis remain unchanged.

The bounds given in Theorem 7.3 give rise to different (weaker) bounds that are stated not in terms of the *degrees* of the elements, but, rather, in terms of what we call the element *connectivities.* For each $e \in \mathscr{E}$ we denote by $\Gamma(e)$ the set of elements that appear together with e in at least one species. More precisely,

$$\Gamma(e) = \{e' \in \mathscr{E} : \operatorname{supp} M_{e'} \cap \operatorname{supp} M_e \neq \emptyset\}\,. \tag{7.20}$$

By the *connectivity* of e we mean the number of such elements:

$$\operatorname{conn}(e) := \#(\Gamma(e)). \tag{7.21}$$

From (7.5) and (7.6) it is not difficult to see that

$$\deg(e) \leq \operatorname{conn}(e) M_e^{max}. \tag{7.22}$$

Thus from Theorem 7.3 we have the following.

Corollary 7.7. *Let $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{K}, \mathscr{E}, \mathscr{M}\}$ be an elemented quasi-thermostatic system with positive equilibrium set E. For each $c^* \in E$ and each $s \in \mathscr{S}$,*

$$\Lambda^s(c^*) \geq \max_{e \in \mathscr{E}} \left\{ \frac{M_e^s}{\operatorname{conn}(e) M_e^{max}} \right\} \tag{7.23}$$

and

$$\Lambda(c^*) \geq \min_{s \in \mathscr{S}} \max_{e \in \mathscr{E}} \left\{ \frac{M_e^s}{\operatorname{conn}(e) M_e^{max}} \right\}. \tag{7.24}$$

8. Constructive reaction networks. By a *constructive reaction network* we mean an elemented network with special properties: Its elements can be identified with certain building blocks that appear explicitly as species of the network and from which all other species are constructed by means of the network's reactions. Example 3.3, which is based on network (1.1) in the introduction, is an example of a constructive reaction network.

Definition 8.1. *A* constructive reaction network *is an elemented reaction network $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{E}, \mathscr{M}\}$ such that $\mathscr{E} \subset \mathscr{S}$ and such that, for each $e \in \mathscr{E}$,*

$$M_e = e + \sum_{q \in \mathscr{Q}} M_e^q q, \tag{8.1}$$

where $\mathscr{Q} = \mathscr{S} \setminus \mathscr{E}$, and where the M_e^q take nonnegative integer values. The members of $\mathscr{Q}$ are called the compounds *of the network.*

For constructive networks the following proposition indicates the sense in which the compounds are built from the elements in terms of reactions appearing in the network.

Proposition 8.2. *Let $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{E}, \mathscr{M}\}$ be a constructive reaction network. Then the set*

$$\left\{ q - \sum_{e \in \mathscr{E}} M_e^q e : q \in \mathscr{Q} \right\} \tag{8.2}$$

is a basis for the stoichiometric subspace S.

Proof. Clearly, the vectors of (8.2) are linearly independent and their number is equal to dim S. To see that they are in fact members of S, it is enough to show that each of the vectors in (8.2) is orthogonal to all the vectors of (8.1). Toward this end, note that for each $q' \in \mathscr{Q}$ and each $e' \in \mathscr{E}$,

$$\left(q' - \sum_{e \in \mathscr{E}} M_e^{q'} e\right) \cdot \left(e' + \sum_{q \in \mathscr{Q}} M_{e'}^q q\right) = M_{e'}^{q'} - M_{e'}^{q'} = 0. \qquad \square$$

Remark 8.3. The vectors of (8.2) can be viewed as "reaction vectors" that derive from the "reactions"

$$\left\{\sum_{e \in \mathscr{E}} M_e^q e \to q : q \in \mathscr{Q}\right\}. \tag{8.3}$$

Each of these "reactions" can be regarded as representing the production of a particular compound *solely* from the elements. Of course, these "construction" reactions need not all be "true" reactions of the original network, but, at least in terms of stoichiometry, each of the "true" reactions can be regarded as a linear combination of the "construction" reactions. Similarly, each member of (8.3) can be viewed as an overall reaction—one for each compound—that derives from a linear combination of the "true" reactions. Hence, the "true" reactions can be viewed as machinery for constructing the compounds from the elements.

Remark 8.4. Proposition 8.2 provides a way to distinguish a reaction network that might be constructive (for a suitable choice of elements) from one that cannot be constructive. For example, network (3.2) has a one-dimensional stoichiometric subspace spanned by $Y + Z - X - W$. Thus, there is no partition of the species set into elements and compounds that is consistent with a basis for the stoichiometric subspace of the form (8.2).

We note that for a *constructive* kinetic system the elements themselves are species, and so it makes sense to speak of their sensitivities. Recall that for a constructive system we have, for each pair $e, e' \in \mathscr{E}$, that $M_e^{e'}$ is 0 if $e' \neq e$ and is 1 otherwise. This observation gives rise to the following corollary to Theorem 7.3.

Corollary 8.5. *Let $\{\mathscr{S}, \mathscr{C}, \mathscr{R}, \mathscr{K}, \mathscr{E}, \mathscr{M}\}$ be a constructive quasi-thermostatic system with positive equilibrium set E. For each $c^* \in E$ and each $e \in \mathscr{E}$,*

$$\Lambda^e(c^*) \geq \frac{1}{\deg(e)} \geq \frac{1}{\operatorname{conn}(e) M_e^{max}}. \tag{8.4}$$

Moreover,

$$\Lambda(c^*) \geq \min_{e \in \mathscr{E}} \left\{\frac{1}{\deg(e)}\right\} \geq \min_{e \in \mathscr{E}} \left\{\frac{1}{\operatorname{conn}(e) M_e^{max}}\right\}. \tag{8.5}$$

Remark 8.6. For any elemented network, it is apparent that, for each element $e \in \mathscr{E}$, $\operatorname{conn}(e) \leq \#(\mathscr{E})$. If we let

$$M^{max} := \max_{e \in \mathscr{E}} \{M_e^{max}\}, \tag{8.6}$$

then we have

$$\deg(e) \leq \operatorname{conn}(e) M_e^{max} \leq \#(\mathscr{E}) M^{max}. \tag{8.7}$$

For a constructive kinetic system in particular we get a crude bound on the system sensitivity:

$$\Lambda(c^*) \geq \frac{1}{\#(\mathscr{E})M^{max}}. \tag{8.8}$$

Although this bound is far less nuanced than (8.5), it already tells us that high system robustness (low sensitivity) requires that the system have a large number of elements or that it contain at least one species having high element content (or a combination of both).

9. Concluding remarks.

Remark 9.1. We emphasized both in the introduction and in section 7 that the species sensitivity bounds given in Theorem 7.3 depend on network structure alone, independent of kinetics or even of the equilibrium state at which they are evaluated. In fact, even the fine details of the reaction network are of limited consequence:

For a quasi-thermostatic kinetic system the entire equilibrium set is determined by specification of just one equilibrium state and the stoichiometric subspace for the underlying network of chemical reactions. This is to say that two elemented quasi-thermostatic systems that share a common equilibrium state and the same stoichiometric subspace have precisely the same positive equilibrium sets and, therefore, *the same species sensitivities (relative to the same choice of elements).* To the extent that the network influences the equilibrium set, then, that influence is felt through the stoichiometric subspace. In turn, the stoichiometric subspace is merely the span of the network's reaction vectors, but the same span might derive from two rather different reaction networks.

Certainly, though, there is *some* network information carried by the stoichiometric subspace S, and therefore, by its orthogonal complement $S^\perp$. Indeed, the sensitivity formulas given by Theorem 7.3 derive squarely from the nature of $S^\perp$ and the particular elemental basis chosen for it.

Remark 9.2. As we indicated earlier, the sensitivity bounds derived for elemented systems suggest that strong robustness (i.e., very low sensitivity) against fluctuations in the element concentrations requires that the elements manifest themselves within the species in high multiplicity or that the species combine with each other gregariously. There is a certain intuitive sense to these requirements, for when they are met, the effect of changes in an element's total concentration can be attenuated through propagation across many species or buffered within species containing multiple instances of that element. It is striking that the requirements for high robustness are highly suggestive of biochemistry, in which large molecules are often built from multiple copies of many distinct elements that readily combine with each other by means of relatively large reaction networks.

Acknowledgments. We thank Noam Berger, Erez Dekel, Yuval Hart, Yakar Kannai, and Ron Milo for their help.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds and Tensor Analysis*, 2nd ed., Springer-Verlag, New York, 1983.

[2] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks:* I. *The injectivity property*, SIAM J. Appl. Math., 65 (2005), pp. 1526–1546.

[3] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks:* II. *The species-reaction graph*, SIAM J. Appl. Math., 66 (2006), pp. 1321–1338.

[4] G. Craciun, Y. Tang, and M. Feinberg, *Understanding bistability in complex enzyme-driven reaction networks*, Proc. Nat. Acad. Sci. USA, 103 (2006), pp. 8697–702.

[5] M. Feinberg, *Complex balancing in general kinetic systems*, Arch. Ration. Mech. Anal., 49 (1972), pp. 187–194.

[6] M. Feinberg, *Mathematical aspects of mass action kinetics*, in Chemical Reactor Theory: A Review, N. Amundsen and L. Lapidus, eds., Prentice–Hall, Englewood Cliffs, NJ, 1977, pp. 1–78.

[7] M. Feinberg, *Lectures on Chemical Reaction Networks*, written version of lectures given at the Mathematical Research Center, University of Wisconsin, Madison, WI, 1979. Available online at http://www.chbmeng.ohio-state.edu/~feinberg/research/.

[8] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors*–I. *The deficiency zero and deficiency one theorems*, Chem. Engrg. Sci., 42 (1987), pp. 2229–2268.

[9] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors*–II. *Multiple steady states for networks of deficiency one*, Chem. Engrg. Sci., 43 (1988), pp. 1–25.

[10] M. Feinberg, *Necessary and sufficient conditions for detailed balancing in mass action systems of arbitrary complexity*, Chem. Engrg. Sci., 44 (1989), pp. 1819–1827.

[11] M. Feinberg, *The existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Ration. Mech. Anal., 132 (1995), pp. 311–370.

[12] V. Guillemin and A. Pollack, *Differential Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[13] R. Heinrich and T. A. Rapoport, *A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector*, European J. Biochem./FEBS, 42 (1974), pp. 97–105.

[14] R. Heinrich and T. A. Rapoport, *A linear steady-state treatment of enzymatic chains. General properties, control and effector strength*, European J. Biochem./FEBS, 42 (1974), pp. 89–95.

[15] J. Higgins, *Analysis of sequential reactions*, Ann. New York Acad. Sci., 108 (1963), pp. 305–321.

[16] F. Horn and R. Jackson, *General mass action kinetics*, Arch. Ration. Mech. Anal., 47 (1972), pp. 81–116.

[17] F. Horn, *Necessary and sufficient conditions for complex balancing in chemical kinetics*, Arch. Ration. Mech. Anal., 49 (1972), pp. 172–186.

[18] K. Israel, *Monotone Behavior for Equilibria of Dynamical Systems*, Ph.D. thesis, Duke University, Durham, NC, 1984.

[19] H. Kacser and J. A. Burns, *The control of flux*, Symposia of the Society for Experimental Biology, 27 (1973), pp. 65–104.

[20] G. Lewis, *A new principle of equilibrium*, Proc. Nat. Acad. Sci. USA, 11 (1925), pp. 179–183.

[21] J. Nailor, *Behavior of Equilibria in Quasi-Thermodynamic Chemical Reaction Networks with Mass-Action Kinetics*, Ph.D. thesis, Duke University, Durham, NC, 1991.

[22] M. Ptashne and A. Gann, *Genes and Signals*, 1st ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.

[23] M. A. Savageau, *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, New York, 1976.

[24] G. Shinar, R. Milo, M. Rodríguez Martínez, and U. Alon, *Input-output robustness in simple bacterial signaling systems*, Proc. Nat. Acad. Sci. USA, 104 (2007), pp. 19931–19935.

[25] R. Veitia, *Nonlinear effects in macromolecular assembly and dosage sensitivity*, J. Theoret. Biol., 220 (2003), pp. 19–25.

[26] R. Wegscheider, *Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reaktionskinetik homogener Systeme*, Z. Phys. Chem., 39 (1902), pp. 257–303.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 999–1023

ANALYSIS OF HEPATITIS C VIRUS INFECTION MODELS WITH HEPATOCYTE HOMEOSTASIS[*]

TIMOTHY C. RELUGA[†], HAREL DAHARI[§], AND ALAN S. PERELSON[¶]

Abstract. Recently, we developed a model for hepatitis C virus (HCV) infection that explicitly includes proliferation of infected and uninfected hepatocytes. The model predictions agree with a large body of experimental observations on the kinetics of HCV RNA change during acute infection, under antiviral therapy, and after the cessation of therapy. Here we mathematically analyze and characterize both the steady state and dynamical behavior of this model. The analyses presented here not only are important for HCV infection but also should be relevant for modeling other infections with hepatotropic viruses, such as hepatitis B virus.

Key words. HCV, viral dynamics, bifurcation analysis

AMS subject classification. 92B99

DOI. 10.1137/080714579

1. Introduction.

Approximately 200 million people worldwide [38] are persistently infected with the hepatitis C virus (HCV) and are at risk of developing chronic liver disease, cirrhosis, and hepatocellular carcinoma. HCV infection therefore represents a significant global public health problem. HCV establishes chronic hepatitis in 60%–80% of infected adults [46]. A vaccine against infection with HCV does not exist, and standard treatment with interferon-α and ribavirin has produced sustained virological response rates of approximately 50%, with no effective alternative treatment for nonresponders to this treatment protocol [13, 30].

A model of human immunodeficiency virus infection [40, 52] was adapted by Neumann et al. [37] to study the kinetics of chronic HCV infection during treatment. Since then viral kinetics modeling has played an important role in the analysis of HCV RNA decay during antiviral therapy (see Perelson [41], Perelson et al. [42]). The original Neumann et al. model for HCV infection [37] included three differential equations representing the populations of target cells, productively infected cells, and virus (Figure 1). A simplified version of the model, which assumes a constant population of target cells, was used to estimate the rates of viral clearance and infected cell loss by fitting to the model the decline of HCV RNA observed in patients during the first 14 days of therapy [37]. However, this simplified version of the model is not able to explain some observed HCV RNA kinetic profiles under treatment [4]. To model complex HCV kinetics, the assumption of a constant level of target cells needs to be relaxed, requiring one to model as correctly as possible the dynamics of the target cell population. Since it has been suggested that hepatocytes, the major cell type in the liver, are also the major producers of HCV [10, 3, 43], we assume here that the

[*] Received by the editors December 9, 2007; accepted for publication (in revised form) September 11, 2008; published electronically January 14, 2009. This work was supported by the U.S. Department of Energy under contract DE-AC52-06NA25396; by NIH grants AI28433, RR06555, AI06525, and P20-RR18754; and by the Human Frontiers Science Program grant RPG0010/2004.

http://www.siam.org/journals/siap/69-4/71457.html

[†] Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (timothy@reluga.org).

[§] Department of Medicine, University of Illinois at Chicago, Chicago, IL 60612 (dahariharihe@uic.edu).

[¶] Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 (asp@lanl.gov).

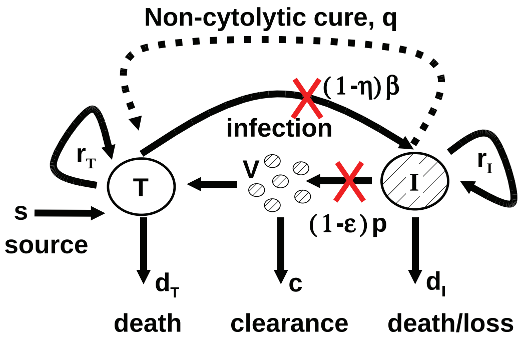


FIG. 1. *Schematic representation of HCV infection models. T and I represent target and infected cells, respectively, and V represents free virus. The parameters shown in the figure are defined in the text. The original model of Neumann et al.* [37] *assumed that there is no proliferation of target and infected cells (i.e., $r_T = r_I = 0$) and no spontaneous cure (i.e., $q = 0$). The extended model of Dahari, Ribeiro, and Perelson* [6]*, which was used for predicting complex HCV kinetics under therapy, includes target and infected cell proliferation without cure ($r_T, r_I > 0$ and $q = 0$). A model including both proliferation and the spontaneous cure of infected cells (dashed line; $q > 0$) was used to explain the kinetics of HCV in primary infection in chimpanzees* [5].

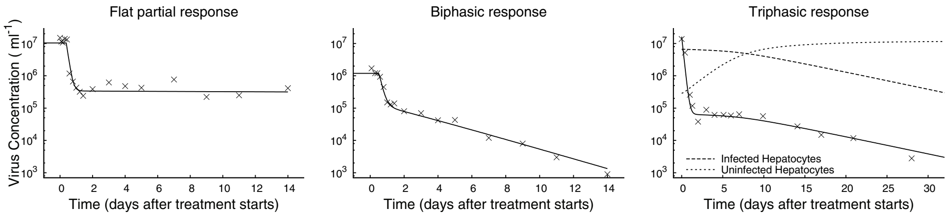
target cells of the model are hepatocytes.

The liver is an organ that regenerates, and due to homeostatic mechanisms, any loss of hepatocytes would be compensated for by the proliferation of existing hepatocytes [12, 32]. However, besides replication of existing hepatocytes, another mechanism of liver cell generation is present (termed here immigration), i.e., differentiation of liver progenitor cells or bone marrow cells [12].

In prior work, we have shown that including proliferation of both target cells and infected cells increases the ability of the model to explain experimental data [4, 6]. Because HCV infection is generally thought to be noncytopathic, i.e., infection per se does not kill a cell [31], proliferation of infected cells has been included in the model. Studying the effects of varying the rate of infected cell proliferation from zero (no proliferation) to rates in excess of the rate proliferation of uninfected cells [34], as might occur by an oncogenic effect, is one of the goals of this work.

HCV is an RNA virus that replicates in the cytoplasm of an infected cell [25]. Due to the action of endogenous nucleases or microRNAs, it is in principle possible for a cell to clear viral RNA [2, 39]. Some of our prior modeling of acute HCV infection in chimpanzees required the inclusion of this type of "cure" of infected cells in order to explain the kinetics of HCV clearance without a massive loss of liver cells that would have led to the animals' death [5]. Thus, the effects of cure of infected cells is also studied in the analysis provided below.

During antiviral therapy for HCV infection, patients may exhibit a flat partial response or a biphasic decline in HCV RNA (Figure 2(left) and (middle)). In addition, a triphasic pattern of HCV RNA decline (Figure 2(right)) has been observed in some treated patients [19]. In these patients, HCV RNA initially falls very rapidly, by 1–2 orders of magnitude during the first day or two of therapy. Then HCV RNA decline ceases, and a "shoulder phase" that can last from days to many weeks is observed. This shoulder phase can persist, in which case it has been called a flat second phase, or it can be followed by a renewed phase of HCV RNA decline; in this case the pattern


Flat partial response

Biphasic response

Triphasic response

FIG. 2. *Three example plots of observed changes in viral load (X) following the start of treatment, together with numerical solutions to system* (2.1) *(solid line). The initial condition of each numerical solution is the chronic-infection steady state. In some cases, there is a flat partial response to treatment (left), where viral load shows an immediate drop but then remains unchanged over time. In some cases, there is a biphasic response (middle), with a rapid initial drop and a slower asymptotic clearance. In some cases, there is a triphasic response, with a rapid initial drop, an intermediate shoulder phase during which there is little change, and then an asymptotic clearance phase. The initial rapid decline in virus load is the synchronization to the new quasi-steady state, following the start of treatment. Afterward, virus load closely tracks the number of infected cells (right). Treatment efficacies are* $\epsilon = 0.98, \eta = 0$ *(left),* $\epsilon = 0.9, \eta = 0$ *(middle), and* $\epsilon = 0.996, \eta = 0$ *(right). Other parameter values are shown in Table* 2.1*.*

has been called triphasic [4, 6]. Another of the goals of this paper is to understand the origin of the triphasic response and to compute from the model the length of the shoulder phase as a function of model parameters. As the length of the shoulder phase approaches zero, a triphasic response transforms into a biphasic response, and thus studying triphasic responses provides a general framework for understanding treatment response kinetics.

In order to accomplish our various goals, we first describe the model and its parameters. Then we study the model's steady states and their stability. Using a perturbation analysis approach, we show how the shoulder phase arises and provide an approximate formula from which one can calculate its length.

2. Model. The model proposed by Dahari and coworkers [6, 4] expands on the standard HCV viral-dynamic model [37] of infection and clearance by incorporating density-dependent proliferation and death (Figure 1). Uninfected hepatocytes, T, are infected at a rate β per free virus per hepatocyte. Infected cells, I, produce free virus at rate p per cell but also die with rate d_I. Free virus is cleared at rate c by immune and other degradation processes. Besides infection processes, hepatocyte numbers are influenced by homeostatic processes. Uninfected hepatocytes die at rate d_T. Both infected and uninfected hepatocytes proliferate logistically with maximum rates r_I and r_T, respectively, as long as the total number of hepatocytes is less than $T_{\max}$. Besides proliferation, uninfected hepatocytes may increase in number through immigration or differentiation of hepatocyte precursors that develop into hepatocytes at constitutive rate $\hat{s}$, or by spontaneous cure of infected hepatocytes through a noncytolytic process at rate $\hat{q}$. Treatment with antiviral drugs reduces the infection rate by a fraction η and the viral production rate by a fraction ϵ. The corresponding system of differential equations is

$$\frac{\mathrm{d}T}{\mathrm{d}\hat{\mathrm{t}}} = \hat{s} + r_T T \left(1 - \frac{T+I}{T_{\max}}\right) - d_T T - (1-\eta)\beta V T + \hat{q} I, \tag{2.1a}$$

$$\frac{\mathrm{d}I}{\mathrm{d}\hat{\mathrm{t}}} = r_I I \left(1 - \frac{T+I}{T_{\max}}\right) + (1-\eta)\beta V T - d_I I - \hat{q} I, \tag{2.1b}$$

$$\frac{\mathrm{d}V}{\mathrm{d}\hat{\mathrm{t}}} = (1-\epsilon) p I - cV, \tag{2.1c}$$

Table 2.1

Estimated parameter ranges for hepatitis C when modeled with system (2.1). *The columns labelled left, middle, and right give the parameter values for fitting system* (2.1) *to the data for the respective plots in Figure* 2. *The* r_T, $T_{\max}$, *and* d_T *parameters are not independently identifiable, so common practice is to fix* d_T *prior to fitting.*

Symbol	Minimum	Maximum	Units	Left	Middle	Right	Ref.
β	10^{-8}	10^{-6}	virus^{-1} ml day^{-1}	1.4×10^{-6}	9.0×10^{-8}	2.8×10^{-8}	[5]
$T_{\max}$	4×10^6	1.3×10^7	cells ml^{-1}	5×10^6	5×10^6	1.2×10^7	[27, 49]
p	0.1	44	virus cell^{-1} day^{-1}	28.7	10.9	13.2	[5]
$\hat{s}$	1	1.8×10^5	cells ml^{-1} day^{-1}	1	1	1	[50]
$\hat{q}$	0	1	day^{-1}	0	0	0	[5]
c	0.8	22	day^{-1}	6.0	5.8	5.4	[45]
d_T	10^{-3}	1.4×10^{-2}	day^{-1}	1.2×10^{-2}	1.2×10^{-2}	1.2×10^{-2}	[26, 28]
d_I	10^{-3}	0.5	day^{-1}	0.36	0.48	0.13	[45]
r_T	2×10^{-3}	3.4	day^{-1}	3.0	0.70	1.1	[5]
r_I	Unknown	Unknown	day^{-1}	.97	0.112	0.26	

where the time $\hat{t}$ is measured in days. Table 2.1 shows estimated ranges for the parameters.

System (2.1) has a three-dimensional phase-space and a twelve-dimensional parameter space, so despite the relative simplicity, the full dynamics are difficult to classify. Fortunately, there are some natural simplifications. The range of rates of viral clearance shown in Table 2.1 is significantly faster than the other time-scale parameters. In numerical simulations (Figure 2), after an initial transient the viral dynamics closely track the dynamics of infected cells. This suggests that viral dynamics can be decomposed into two time scales: a fast time scale starting at $\hat{t}_0$ where the number of infected hepatocytes, I, is relatively constant and

$$V(\hat{t}) \approx \frac{(1-\epsilon)p}{c} I(\hat{t}_0) + \left[V(\hat{t}_0) - \frac{(1-\epsilon)p}{c} I(\hat{t}_0) \right] e^{-c(\hat{t}-\hat{t}_0)}, \tag{2.2}$$

and a slow time scale where

$$V(\hat{t}) \approx \frac{(1-\epsilon)p}{c} I(\hat{t}). \tag{2.3}$$

For patients in steady state before treatment, as is typically the case, $I(\hat{t}_0) = cV(\hat{t}_0)/p$, allowing one to simplify (2.2) to

$$V(\hat{t}) = (1-\epsilon)V(\hat{t}_0) + \epsilon V(\hat{t}_0) e^{-c(\hat{t}-\hat{t}_0)}. \tag{2.4}$$

On time scales longer than $1/c$, then, the dynamics of system (2.1) can be approximated by a system of two equations. If we now introduce the dimensionless time $t = (r_T - d_T)\hat{t}$, the dimensionless state variables

$$x = \frac{T}{T_{\max}}, \quad y = \frac{I}{T_{\max}}, \tag{2.5}$$

and the dimensionless parameters

$$\begin{gathered} s = \frac{\hat{s} r_T}{(r_T - d_T)^2 T_{\max}}, \quad b = \frac{p \beta T_{\max}}{c r_T}, \quad q = \frac{\hat{q}}{r_T - d_T}, \\ r = \frac{r_I}{r_T}, \quad d = \frac{d_I r_T - d_T r_I}{r_T (r_T - d_T)}, \quad 1 - \theta = (1-\epsilon)(1-\eta), \end{gathered} \tag{2.6}$$

then under the quasi-steady state approximation, system (2.1) is equivalent to the dimensionless system

$$\dot{x} = x\,(1 - x - y) - (1 - \theta)byx + qy + s, \tag{2.7a}$$

$$\dot{y} = ry\,(1 - x - y) + (1 - \theta)byx - dy - qy. \tag{2.7b}$$

Note that a fundamental assumption in the transformation to system (2.7) is that $r_T > d_T$, which we expect because of the hepatocyte population's ability to support itself and to regenerate itself after injury.

Immigration of new hepatocytes is believed to be slow ($< 1\%$ per day; Table 2.1) relative to the total number of hepatocytes (i.e., $s \ll 1$). Spontaneous cure from HCV has not yet been directly observed. It has been suggested to occur based on the kinetics of HCV clearance and liver damage in humans [51] and in chimpanzees [5]. Therefore, in a first analysis, we assume that $s = q = 0$. Later, we reintroduce these parameters and examine their effects via a perturbation analysis. Dropping the s and q terms, system (2.7) simplifies to

$$\dot{x} = x\,(1 - x - y) - (1 - \theta)byx, \tag{2.8a}$$

$$\dot{y} = ry\,(1 - x - y) + (1 - \theta)byx - dy. \tag{2.8b}$$

Most of the parameter ranges from Table 2.1 are captured by allowing $b \in [10^{-2}, 10^3]$ and $d \in [10^{-3}, 10^2]$. r_I has not yet been studied experimentally, and thus we cannot bound r beyond the trivial statement that $r \geq 0$.

Gómez-Acevedo and Li [14] have previously studied some of the properties of system (2.8) in the context of human T-cell lymphotropic virus type I. It is a simple model with only three independent parameters and dynamics that can be completely analyzed using phase-plane analysis and algebraic methods while still encapsulating the fundamental concepts of system (2.1). system (2.8) diverges from common viral dynamics models in the homeostasis parameter r. When $r = 0$, system (2.8) is naturally interpreted as an epidemic model, a viral infection model, or a predator-prey model. When the epidemic model is extended to include logistic homeostasis with $r > 0$, the infected cells can also proliferate independent of x but experience additional density-dependent mortality as a function of the total population size $x+y$. This paper explores the consequences of this homeostasis. We will first study systems (2.7) and (2.8) during acute infection. We will then study the response of these systems to treatment.

3. Dynamics without treatment ($\boldsymbol{\theta = 0}$).

When HCV first infects a person, the ensuing dynamics depend on the relative parameter values. Since newly infected individuals do not know that they are infected, we assume that there is initially no treatment ($\theta = 0$). At first glance, we might expect several different scenarios to ensue following exposure: infection may fade out without becoming established, infection may spread with limited success and infect only part of the liver, or infection may spread rapidly and infect the whole liver. To understand when the dynamics of system (2.7) under acute infection correspond to each of these situations, it is helpful to walk through the bifurcation structure of system (2.8).

3.1. Without immigration or spontaneous cure.

When there is no immigration ($s = 0$) or spontaneous cure ($q = 0$), the dynamics are described by system (2.8). system (2.8) is a variant of the Lotka–Volterra equations studied extensively in ecology [21]. The $\dot{x}$-nullclines are $x = 0$ and $y = (1 - x)/(1 + b)$. The $\dot{y}$-nullclines

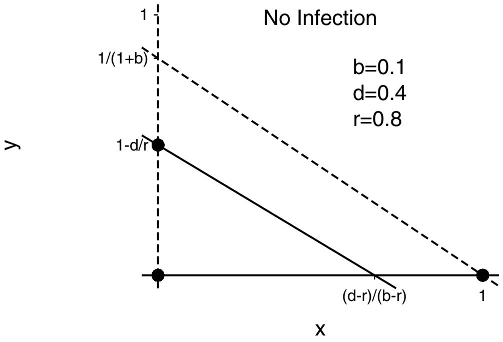


FIG. 3. *Example nullclines for system* (2.8) *when the disease-free equilibrium is globally attracting. The liver-free ($x = y = 0$), disease-free ($x = 1, y = 0$), and total-infection ($x = 0, y = 1 - d/r$) stationary solutions are marked with dots. The solid lines are the $\dot{y}$-nullclines, and the dotted lines are the $\dot{x}$-nullclines. The partial-infection stationary solution is not present for these parameter values.*

TABLE 3.1
Stationary solutions for system (2.8) *and their characteristics.*

Stationary point	Location	Bifurcation conditions	Local stability condition
Liver-free	$(0, 0)$	$r = d$	Never stable
Disease-free	$(1, 0)$	$b = d$	$b < d$
Total-infection	$(0, 1 - d/r)$	$r = d + d/b$, $r = d$	$r > d + d/b$
Partial-infection	$\left(\frac{db + d - br}{b(1 + b - r)}, \frac{b - d}{b(1 + b - r)}\right)$	$r = d + d/b$, $b = d$	$rb/(1 + b) < d < b$

are $y = 0$ and $y = 1 - d/r - (1 - b/r)x$ (see Figure 3). Up to four stationary solutions to system (2.8) can be found at the intersections of the $\dot{x}$ and $\dot{y}$ nullclines. They are

$$(0, 0), \quad (1, 0), \quad \left(0, 1 - \frac{d}{r}\right), \quad \text{and} \quad \left(\frac{db + d - br}{b(1 + b - r)}, \frac{b - d}{b(1 + b - r)}\right), \tag{3.1}$$

respectively the liver-free solution, the disease-free solution, the total-infection solution, and the partial-infection solution. The locations and stability conditions for the stationary solutions are summarized in Table 3.1.

The bifurcations and stability of these four stationary solutions depend on the parameter values in ways summarized in Table 3.1. system (2.8)'s Jacobian is

$$\mathbf{J} = \begin{bmatrix} 1 - 2x - y - by & -x(1 + b) \\ -ry + by & r(1 - x - 2y) - d + bx \end{bmatrix}. \tag{3.2}$$

The classification of the parameter space is summarized in Figure 4, with examples of each region's nullclines given in Figure 5. Before treatment, the reproductive number of infection

$$\mathcal{R} = \frac{b}{d} \tag{3.3}$$

at the disease-free equilibrium $(1, 0)$. In order for HCV to infect the liver, $\mathcal{R}$ must be greater than 1, indicating that on average an infected hepatocyte causes more than one uninfected cell to become infected. The eigenvalues at the disease-free equilibrium

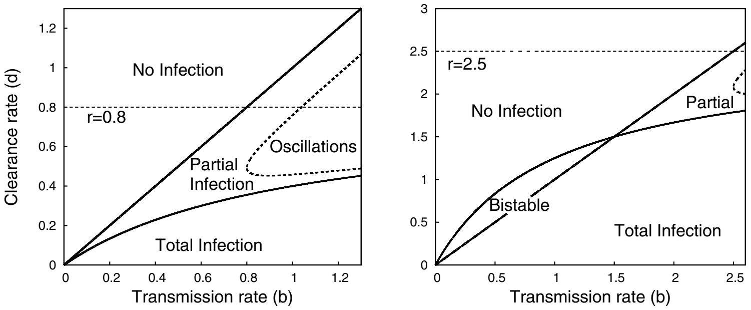


Fig. 4. *Plot representing the parameter regions for asymptotic dynamics of system* (2.8) *when* $r = 0.8$ *(left) or* $r = 2.5$ *(right). Within the region marked partial infection, the dotted line is the boundary between monotone convergence and oscillatory convergence to the partial infection steady state.*

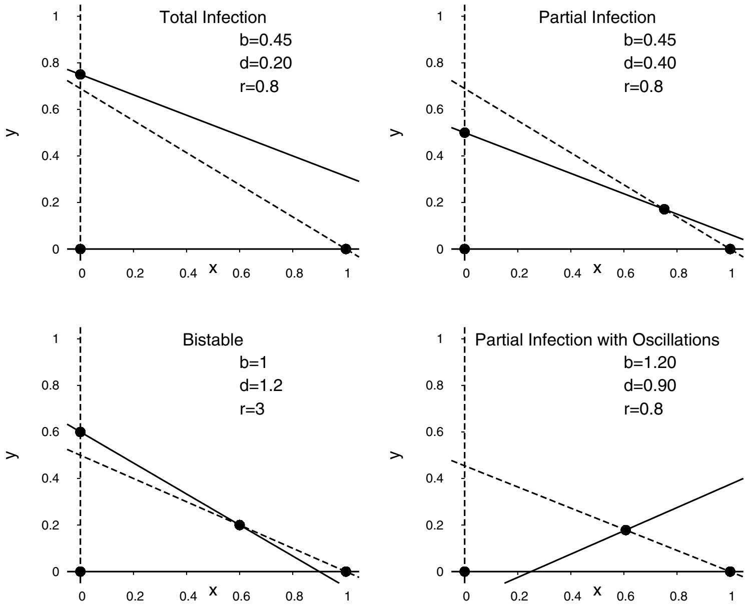


Fig. 5. *Example phase planes of system* (2.8) *for distinct parameter regions. The dashed lines are the* $\dot{x}$*-nullclines, and the solid lines are the* $\dot{y}$*-nullclines. The dots represent stationary solutions.*

are $\lambda = -1$, corresponding to the eigenvector $[1, 0]$, and $\lambda = b - d$, corresponding to the eigenvector $[-1 - b, 1 + b - d]$. If $\mathcal{R} < 1$, the disease-free solution $(1, 0)$ is locally attracting. If $\mathcal{R} > 1$, HCV infects new cells faster than infected cells die, and the asymptotic dynamics may correspond to either partial or total infection of the liver.

The liver-free stationary solution $(0, 0)$ is always unstable, switching between a saddle point when infected cells die quickly ($r < d$) and an unstable node when infected cells die slowly ($d < r$). If the proliferation rate is slower than the death rate ($r < d$), then HCV can never totally infect the liver. There is a transcritical bifurcation at $d = r$, and the total-infection stationary solution $(0, 1 - d/r)$ is feasible only when the proliferation rate of infected cells is greater than the excess death rate

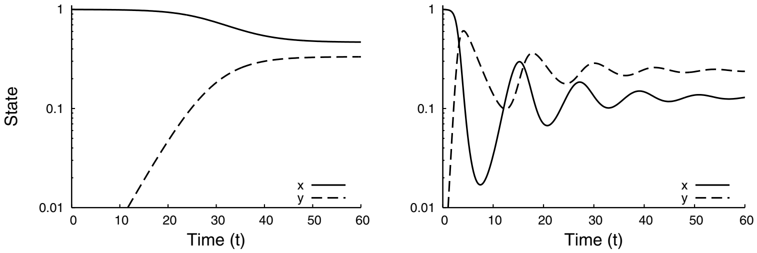


Fig. 6. *Time series for system* (2.7) *of monotone (left,* $r = 0.6, b = 0.6$*) and oscillatory (right,* $r = 0.1, b = 2.6$*) convergence to the partial infection stationary solution when* $d = 0.4$*.*

of infected cells (Figure 4). From the Jacobian, we see that if $d+d/b < r$ (equivalently, $d < rb/(1+b)$), total infection is locally stable, and from the general theory of Lotka–Volterra systems, it is globally stable, provided $d < \min\{b, \frac{rb}{1+b}\}$. This includes all cases where $d < 0$.

The partial-infection stationary solution is present whenever d lies between b and $\frac{rb}{1+b}$. The local stability of the partially infected stationary solution can be determined from the characteristic polynomial

$$\lambda^2 + \frac{d}{b}\lambda + \frac{(d-b)(br-bd-d)}{b(1+b-r)} = 0, \tag{3.4}$$

where λ is an eigenvalue. If $b < d < \frac{rb}{1+b}$, the constant term of the characteristic polynomial at the partial-infection stationary solution is negative, implying (by Decartes' rule of sign) that there is a single positive root and that the partial-infection steady state is a saddle point. In this situation, we can show that both the disease-free and the total-infection stationary solutions are locally stable. As first shown in Gómez-Acevedo and Li [14], the system is bistable, and the asymptotic dynamics will depend on the initial conditions. The constraint $r < 1$ is sufficient to preclude bistability.

When $b > d > \frac{rb}{1+b}$, the coefficients d/b and

$$\frac{(d-b)(br-bd-d)}{b(1+b-r)} \tag{3.5}$$

of the characteristic polynomial are both positive. From the Routh–Hurwitz conditions [35], it follows immediately that the partial-infection stationary solution is locally stable. From prior work on Lotka–Volterra equations [20], we know that it is also globally stable.

Convergence to the partial-infection stationary solution can be oscillatory if the eigenvalues are complex, or monotone if the eigenvalues are real (Figure 6). Calculation of the discriminant shows that the convergence is oscillatory whenever

$$\frac{d^2}{4b^2} - \frac{(b-d)(rb-db-d)}{b(r-1-b)} < 0. \tag{3.6}$$

This inequality is not easy to interpret by inspection, but it is quadratic in d, and so it is easy to handle numerically. The boundaries of the subset of parameter space where convergence is oscillatory asymptotically converge to $d(b) = r$ and $d(b) = b$ as b diverges to ∞. When convergence is oscillatory, the period of oscillations around

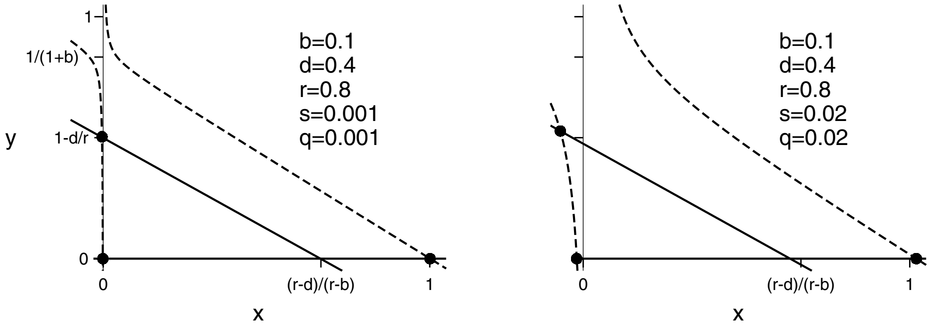


FIG. 7. *Example nullclines for system* (2.7) *when the disease-free equilibrium is globally attracting for immigration and spontaneous clearance rates* $s = q = 0.001$*(left) or* $s = q = 0.02$ *(right). The liver-free, disease-free, and total-infection stationary solutions are marked with dots. The liver-free and total-infection stationary solutions have negative* x *coordinates when the immigration and spontaneous clearance rates are positive. The solid lines are the* $\dot{y}$*-nullclines, and the dotted lines are the* $\dot{x}$*-nullclines. The partial-infection stationary solution is not present for these parameter values.*

the stationary solution is

$$\frac{2\pi}{\sqrt{\frac{d^2}{4b^2} - \frac{(b-d)(rb-db-d)}{b(r-1-b)}}}. \tag{3.7}$$

A sufficient condition for monotone convergence to the partial-infection stationary solution instead of oscillations is $b < r$, in which case the convergence rate is governed by the slowest eigenvalue,

$$-\frac{d}{2b} + \sqrt{\frac{d^2}{4b^2} + \frac{(b-d)\,(b\,r - d - b\,d)}{b\;(b+1-r)}}. \tag{3.8}$$

3.2. With immigration and spontaneous cure. Including immigration ($s > 0$) and spontaneous clearance ($q > 0$) in system (2.7) changes the dynamics of system (2.8) in small but important ways (Figure 7). The two $\dot{y}$-nullclines are $y = 0$ and

$$y = \left(\frac{b}{r} - 1\right)x + 1 - \frac{d+q}{r}. \tag{3.9}$$

Spontaneous clearance moves the nullcline given by (3.9) slightly to the left, but the $\dot{y}$-nullclines are basically the same as those of system (2.8). The change in the $\dot{x}$-nullclines is more pronounced. The only $\dot{x}$-nullcline in system (2.7) is

$$y = \frac{s + x(1-x)}{x + bx - q}. \tag{3.10}$$

The $\dot{x}$-nullclines have changed from a pair of intersecting lines in system (2.8) to a hyperbola in system (2.7). The shape of the hyperbola is still the same as those of system (2.8) except near the intersection point $(0, 1 - d/r)$. The hyperbola is also shifted slightly down and to the right compared to system (2.8) (Figure 7). For large positive and negative x, the nullcline is approximately equal to $(1-x)/(1+b)$. There

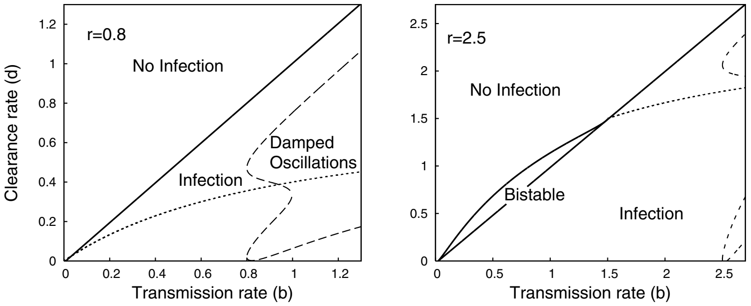


Fig. 8. *Plot representing the parameter regions for asymptotic dynamics of system* (2.7) *with* $s = 0.01$, $q = 0$ *when* $r = 0.8$ *(left) or* $r = 2.5$ *(right). Compare to Figure* 3. *The dashed line represents the boundary of the parameter region where convergence to the steady state exhibits damped oscillations. The dotted line represents the bifurcation boundary between partial and total infection when* $s = q = 0$. *However, there is no formal bifurcation between partial and total infection if* s *or* q *is positive because of the structural instability of the transcritical bifurcation in system* (2.8).

is a vertical asymptote at $x = q/(1 + b)$. The nullcline is positive just to the right of this asymptote and negative just to its left. The $\dot{x}$-nullcline's unique y-intercept is $y = -s/q$. This implies that there can be no biologically feasible stationary solutions with $x \leq q/(1 + b)$; i.e., total hepatocyte loss is no longer a stationary solution because the model now includes a perpetual source of new hepatocytes. This change also means that there is no longer a bifurcation between partial and total infection (see Figure 8).

Two stationary solutions to system (2.7) solve

$$x^2 - x - s = 0 \quad \text{with} \quad y = 0. \tag{3.11}$$

The exact solutions are

$$\left(\frac{1 \pm \sqrt{1 + 4s}}{2}, 0\right). \tag{3.12}$$

When s is very small, the solutions of (3.11) are approximately

$$(-s + o(s), 0) \quad \text{and} \quad (1 + s + o(s), 0). \tag{3.13}$$

The solution with the negative square root can never appear biologically because it predicts a negative number of uninfected hepatocytes.

The other two stationary solutions of system (2.7) solve

$$b\,(1 + b - r)\,x^2 + [(rb - db - d) + q\,(r - 1 - 2\,b)]\,x - sr - (r - d - q)\,q = 0, \tag{3.14a}$$

$$\text{with} \quad y = \left(\frac{b}{r} - 1\right) x + 1 - \frac{d + q}{r}. \tag{3.14b}$$

The solutions can be expressed in terms of radicals, but greater intuition of the effects of s and q relative to the stationary solutions of system (2.8) can be gained through perturbation analysis (see Appendix A). When $d < \min\{b, rb/(1 + b)\}$, the one biologically meaningful solution to (3.14) is

(3.15)

$$(x,y) = \left(\frac{q(r-d)+sr}{rb-bd-d} + o(s,q), 1 - \frac{d}{r} + \frac{(r-b)\,s}{d+bd-br} - \frac{\left(r^2-rd-d\right)q}{r\,(-d-bd+br)} + o(s,q) \right),$$

corresponding to the total-infection stationary solution of system (2.8) but with a small number of uninfected cells sustained by the sources s and q. The $o(s,q)$ terms in (3.16) hide higher order effects in s and q that vanish quadratically or faster as s and q approach zero. When $rb/(1+b) < d < b$,

$$x = \frac{d+db-rb}{b\,(1+b-r)} - \frac{rs}{rb-d-db} + \frac{\left(rb^2+rd-d-2db-db^2\right)}{b\,(1+b-r)\,(rb-d-db)}q + o(s,q), \tag{3.16a}$$

$$y = \frac{b-d}{b\,(1+b-r)} - \frac{s\,(b-r)}{rb-d-db} - \frac{\left(b^2-2\,db+rd-d\right)q}{b\,(1+b-r)\,(rb-d-db)} + o(s,q) \tag{3.16b}$$

is an approximate solution corresponding to the partial-infection stationary solution of system (2.8). The other solution of (3.14) is negative.

When $b < d < rb/(1+b)$, both solutions to (3.14) are positive. Again, this can occur only when $r > 1$, i.e., when infected cells proliferate faster than uninfected ones. Approximate locations are given by (3.15) and (3.16). The bifurcation between zero and two roots is a saddle-node bifurcation where the root with smaller x value is a stable node and the root with larger x value is a saddle. The calculation of the exact condition for bistability when s or q is positive is algebraically opaque, requiring the solution of a pair of polynomials that are quadratic in d and a test to distinguish bistability outside the positive quadrant from bistability inside the positive quadrant. The net effect in system (2.7) of this complexity is a minor perturbation of that found for system (2.8) (compare Figures 4 and 8). Immigration and spontaneous clearance shrink the bistable region of parameter space slightly and shift it so that it occurs for slightly smaller values of d.

The local stability of the stationary solutions to system (2.7) is predicted by the Jacobian matrix

$$\mathbf{J} = \begin{bmatrix} 1-2x-y-by & -x(1+b)+q \\ -ry+by & r\,(1-x-2y)-d+bx-q \end{bmatrix}. \tag{3.17}$$

The disease-free stationary solution loses stability through a transcritical bifurcation that occurs at $\det \mathbf{J} = 0$. Substituting $y = 0$ into $\mathbf{J}$, $\det \mathbf{J} = 0$ if $x = 1/2$ or $(b-r)x = d+q-r$. Using the approximation $x = 1+s+o(s)$, we can show that the disease-free stationary solution is stable when

$$(1+s)b < d+q+rs+o(s). \tag{3.18}$$

Local stability of the other stationary solutions to system (2.7) can also be approximated analytically, but resulting formulas are difficult to interpret.

4. Treatment effects.

Treatment effects appear in systems (2.7) and (2.8) only through a multiplicative factor $(1-\theta)$ reducing the transmission rate b, where θ is the dimensionless treatment efficacy. Thus, the stationary solution structure of systems (2.7) and (2.8) under treatment is summarized by replacing the x-axis labels

Table 4.1

Classification of the dynamics of system (2.7) *for* $b > d$, $0 < s \ll 1$, $0 < q \ll 1$. *These classifications are only approximate. Parameter values that fall near the boundaries of any of these regions may have dynamics that fit multiple classifications. See Figures* 9 *and* 10 *for graphical depictions and example time series.*

Pretreatment	Pretreatment state	Treatment	Treatment dynamics
$r < d$	Partial infection		
		$\theta < \theta_c$	Reduced infection
		$\theta > \theta_c$	Clearance, no treatment delay
$d < r < d + \frac{d}{b}$	Partial infection		
		$\theta < \theta_c$	Reduced infection
		$\theta > \theta_c$	Clearance, weak treatment delay
$d + \frac{d}{b} < r < d + 1$	Near total infection		
		$\theta < \theta_p$	No effect
		$\theta_p < \theta < \theta_c$	Reduced infection
		$\theta_c < \theta$	Clearance, strong treatment delay
$d + 1 < r$	Near total infection		
		$\theta < \theta_p$	No effect
		$\theta_p < \theta < \theta_c$	Bistable, no effect
		$\theta_c < \theta$	Clearance, strong treatment delay

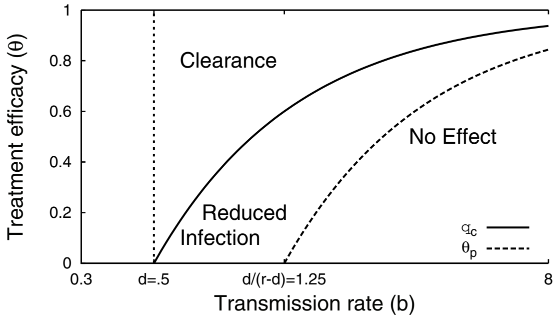


Fig. 9. *The treatment efficacy* θ *leading to specific dynamics for various transmission rates* b, *given* $d = 0.5$, $r = 0.9$, $s = 0$, *and* $q = 0$. *Treatment efficacies below* θ_p *have little or no effect on the number of infected hepatocytes. Treatment efficacies between* θ_c *and* θ_p *reduce the number of infected hepatocytes but are not sufficient for complete clearance. Treatment efficacies greater than* θ_c *lead to complete clearance of infection.*

in Figures 4 and 8 by $(1-\theta)b$. Taking b to be constant, the outcome of drug treatment depends on the drug efficacy θ (see Table 4.1 and Figure 9). There is a critical efficacy

$$\theta_c \approx \begin{cases} 1 - \frac{d+q+rs}{(1+s)b} & \text{if} \quad r < d+1, \\ 1 - \frac{d}{(r-d)b} & \text{if} \quad r > d+1 \end{cases} \tag{4.1}$$

such that $\theta > \theta_c$ implies that treatment will clear the infection. When $r < d + 1$, the critical efficacy θ_c corresponds to reducing the reproductive number to 1. When $r > d + 1$, the treatment efficacy must be large enough not only to reduce the reproductive number below 1, but also to overcome the local stability of the total-infection stationary solution in the region of bistability.

Below the critical efficacy θ_c, there is also a fuzzy partial-efficacy threshold

$$\theta_p \approx \begin{cases} 1 - \frac{d+q+rs}{(1+s)b} & \text{if} \quad r > d+1, \\ 1 - \frac{d}{(r-d)b} & \text{if} \quad d + d/b < r < d+1, \\ 0 & \text{if} \quad r < d + d/b. \end{cases} \tag{4.2}$$

The partial-efficacy threshold was derived using the approximate disease-free stability condition $(1-\theta)b(1+s) < d+q+rs$ and the approximate bifurcation condition $d = r(1-\theta)b/(1+(1-\theta)b)$ between partial and total infection. These conditions were chosen to correspond approximately to those of system (2.8).

A fuzzy threshold is a weaker form of the standard threshold conditions used in bifurcation analysis. Thresholds like the critical efficacy threshold θ_c indicate the location of a bifurcation or discontinuity of some form. On a given side of the threshold, changes in parameters can be interpreted as continuous smooth changes in the system. On opposite sides of the threshold, dynamics are qualitatively distinct, and parameter changes that cross the threshold typically cause nonsmooth and discontinuous changes in the system. But in many systems, important differences in dynamics are not separated by a discontinuity; the system can change in a continuous, smooth manner between qualitatively different extremes. Since there is no discontinuity in the system, we cannot define an exact threshold. As a next best recourse, we define a fuzzy threshold that in some sense separates regions of parameter space with different dynamics. Unlike standard thresholds, which are uniquely defined by their discontinuity, fuzzy thresholds are not uniquely defined; there are infinitely many fuzzy thresholds that distinguish well-separated points in parameter space, and points in the neighborhood of one fuzzy threshold may be classified in any variety of ways by other fuzzy thresholds. Still, fuzzy thresholds can serve as useful rules of thumb. θ_p is a fuzzy threshold because the transition between total infection and partial infection in system (2.7) does not generally coincide with a bifurcation.

Treatment efficacies between θ_p and θ_c can significantly reduce the fraction of hepatocytes that are infected but will not clear infection completely. Treatment efficacies $\theta < \theta_p$ do not significantly reduce the fraction of hepatocytes infected (Figure 9).

An important aspect of treatment response is the dynamics of the transition from the pretreatment state to the posttreatment stationary solution. When treatment is only partially effective ($\theta_p < \theta < \theta_c$), the dynamics converge to a new partial-infection stationary solution, and this convergence can be monotone, can overshoot, and can show damped oscillations, depending on the parameter values. When treatment is above the critical efficacy ($\theta_c < \theta$), infection asymptotically decays at rate $(1-\theta)b - d$. However, in situations with near-total infection, ($r > d + d/b$), there can be a significant delay before the number of infected hepatocytes begins to decay (see Figure 10). This delay may be a "strong" delay, where the number of infected hepatocytes does not change for an extended period of time after the start of treatment before decaying exponentially, or a "weak" delay, where the number of infected hepatocytes decays slowly at the start of treatment but then accelerates (Figure 10). The role of the relative proliferation rate r in treatment response is shown in Figure 11. These strong and weak delays are important because they may correspond to the "shoulder phase" observed in HCV viral load time series after the start of treatment [6].

Why does this delay occur? Suppose that treatment is highly effective ($\theta > \theta_c$). Biologically, treatment shifts the competitive advantage away from infected to uninfected hepatocytes, but there are initially too few uninfected hepatocytes to displace a

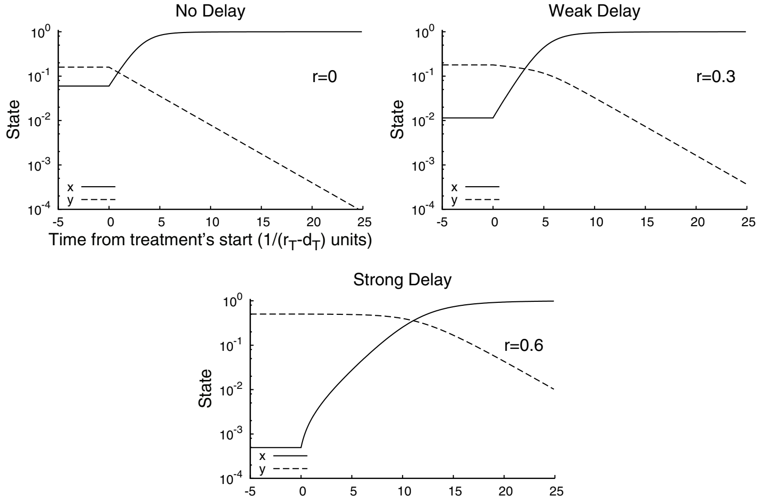


FIG. 10. *Time series for system* (2.7) *with treatment starting at* $t = 0$ *for* $r = 0$ *(top left),* $r = 0.3$ *(top right), and* $r = 0.6$ *(bottom). The initial condition is the pretreatment stationary solution. When the proliferation rate* r *is small (top left), there is no delay; the number of infected cells (y) decays at a constant rate from the start of treatment. For intermediate proliferation rates (top right), there is a weak delay between the start of treatment and the asymptotic clearance of infection. When* r *is large (bottom), there is a strong delay (about* 10 *units here) before the number of uninfected cells (x) reaches equality with the number of infected cells and the decay rate of infected cells accelerates to its exponential asymptotic rate. Parameter values* $s = 0.001, q = 0, d = 0.3, b = 5, \theta = 1$.

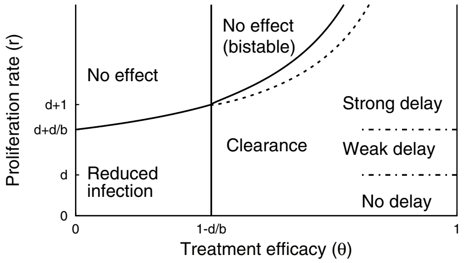


FIG. 11. *Classification of treatment-response as a function of* θ *and* r *when* $b > d$*;* $q = 0$*,* $s = 0.001$*,* $b = 0.9$*,* $d = 0.5$*. Regions are labeled according to the dynamics observed under treatment, assuming that the dynamics were at equilibrium prior to treatment. In the bistable region, both the disease-free and total-infection stationary solutions are locally stable under treatment. The boundaries between the regions of strong delay, weak delay, and no delay are fuzzy in the case of* $\theta = 1$*, and the boundaries are even fuzzier for* $\theta < 1$*. In the sliver between the dotted line and the solid line defining the bistable region our approximation to* t_d *in* (4.8) *fails because there is no nearby stationary solution to use for* $\mathbf{u}^*$ *(see Figure* 12*). In this sliver, the approximation method described in Appendix* C *can be used.*

significant portion of the infected hepatocytes. Infected hepatocytes begin to decline only when the number of uninfected hepatocytes reaches the same order of magnitude.

The rate of recovery depends on some details of the phase-plane geometry. Let's consider how the phase plane changes as we increase the efficacy θ when $r > d + 1$. While $\theta < \theta_p$ the total-infection stationary solution is attracting and the disease-free stationary solution is a saddle point. As θ increases to between θ_p and θ_c, the

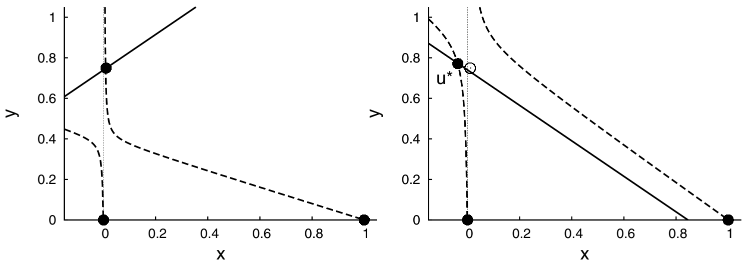


FIG. 12. *Nullclines of system* (2.7) *when* $s = 0.001$*,* $q = 0.008$*,* $r = 0.8$*,* $d = 0.2$*, and* $b = 1.5$ *with before treatment,* $\theta = 0$ *(left) and at the start of treatment,* $\theta = 14/15$ *(right). The solid dots represent stationary solutions. The open dot in the right-hand plot corresponds to the attracting stationary solution in the left-hand plot and is the initial condition for the dynamics when treatment begins. The adjacent solid dot* $\mathbf{u}^*$ *is the unstable stationary solution around which we linearize to approximate the treatment delay.*

disease-free stationary solution becomes locally stable, but the total-infection stationary solution is still stable and attracts orbits from the pretreatment initial condition. As θ is increased just beyond $\theta + c$, the total-infection stationary solution collides in a saddle-node bifurcation with the unstable partial-infection stationary solution, and both solutions disappear. Now, infection will be cleared by treatment. The rate of clearance is controlled by a bottleneck left in the region of the phase-plane where the saddle-node bifurcation occurred (see Appendix C). As the efficacy θ increases further, the bottleneck weakens, but another saddle-node bifurcation occurs in the second quadrant of the phase-plane. The saddle-node bifurcation introduces a new saddle-point stationary solution close to the pretreatment initial condition. As the bottleneck from the first saddle-node bifurcation is relaxed, the unstable manifold of the new saddle point becomes the primary factor controlling the recovery rate of uninfected hepatocytes.

The distinctions between a strong delay, a weak delay, and no delay (Figures 10 and 11) are empirically determined ones. When treatment completely prevents new infections ($\theta = 1$), we observe in numerical solutions strong delays when $r > d + d/b$, weak delays when $d + d/b > r > d$, but no delays when $r < d$. However, these are only observational distinctions, and the classification of delays is less clear for less efficient treatments.

The existence of a treatment delay is most clean-cut in cases like that of Figure 12, where almost all hepatocytes are infected before treatment and treatment is highly effective. We will now describe a method for approximating the dynamics at the start of treatment and determining the delay, t_d, before the number of infected hepatocytes begins to decline in these cases. We can use the linearization of system (2.7) near the new unstable stationary solution $\mathbf{u}^* = (x^*, y^*)$ (Figure 12) given by the solutions of

$$\begin{aligned}(4.3\text{a})\quad 0 = {}&(1-\theta)b\left(1+(1-\theta)b-r\right)(x^*)^2\\ &+\left[(r(1-\theta)b-d(1-\theta)b-d)+q\left(r-1-2\left(1-\theta\right)b\right)\right]x^*-sr-\left(r-d-q\right)q\end{aligned}$$

$$(4.3\text{b})\qquad \text{with}\quad y^* = \left(\frac{(1-\theta)b}{r}-1\right)x^*+1-\frac{d+q}{r}$$

that is nearest to the positive quadrant. In the neighborhood of $\mathbf{u}^*$, the solution of system (2.7) is approximately given by

$$\mathbf{u}(t) = \mathbf{u}^* + e^{\mathbf{J}(\mathbf{u}^*)t}\left(\mathbf{u}(0) - \mathbf{u}^*\right), \tag{4.4}$$

where $\mathbf{u}(0)$ is the pretreatment equilibrium and $\mathbf{J}(\mathbf{u}^*)$ is the Jacobian matrix at $\mathbf{u}^*$. The matrix exponential can be conveniently expressed in terms of the Lagrange interpolation formula [33]:

$$e^{\mathbf{J}t} = \sum_{n=1}^{N} e^{z_n t} \prod_{i \neq n} \left(\frac{\mathbf{J} - z_i \mathbf{I}}{z_n - z_i} \right), \tag{4.5}$$

where z_n is the nth eigenvalue and $\mathbf{I}$ is the identity matrix. When s and q are small, $\mathbf{u}^* = (x^*, y^*) \approx (0, 1 - d/r)$ (see Appendix A), so the Jacobian

$$\mathbf{J}(\mathbf{u}^*) \approx \begin{bmatrix} \frac{d}{r}\left(1 + (1-\theta)b\right) - (1-\theta)b & 0 \\ \left((1-\theta)b - r\right)\left(1 - \frac{d}{r}\right) & d - r \end{bmatrix}. \tag{4.6}$$

The eigenvalues are approximately

$$\frac{d}{r}\left(1 + (1-\theta)b\right) - (1-\theta)b \qquad \text{and} \qquad d - r. \tag{4.7}$$

Exact formulas can be obtained using the radical expressions for $\mathbf{u}^* = (x^*, y^*)$.

As the final part of the process of determining the treatment delay t_d, we have to identify a condition that marks the end of a treatment delay and agrees with our intuitive observations. There are many possible choices (see Appendix B for a discussion). We found that the condition $x(t_d) = y(t_d)$, corresponding to the point where the number of uninfected cells equals the number of infected cells, was simple, convenient, and robust for calculating t_d over the strong-delay parameter range. Solving for t_d, we find

$$t_d = \frac{r}{d + (1-\theta)b(d-r)} \log \left\{ \frac{[(r - (1-\theta)b)(r-d) + d](y^* - x^*)}{[2(r - (1-\theta)b)(r-d) + d](x(0) - x^*)} \right\}. \tag{4.8}$$

The dimensional delay time is $\hat{t}_d = \frac{t_d}{r_T - d_T}$ days. A side-by-side comparison of (4.8) to the actual value calculated by numerical solution of system (2.7) is shown in Figure 13. The figure shows that our approximation gives results that are very similar to the numerical solutions.

If the relative proliferation rate of infected cells r is fixed at a sufficiently large value ($r - d \geq 1$, for example), then the treatment delay increases as the excess death rate d decreases and the transmission rate b increases (see Figure 14). We see from Figures 15 and 16 that the treatment delay increases as r increases, until r is sufficiently large to introduce bistability, corresponding to $t_d \to \infty$. In the strong-delay region of Figure 11, the larger the immigration rate s of uninfected hepatocytes, the shorter the treatment delay because there are more uninfected hepatocytes at the start of treatment (compare Figures 15 and 16). The effect of spontaneous cure (q) is similar to that of immigration (s); more spontaneous cure shortens the treatment delay (Figure 17). The sensitivity to immigration and spontaneous cure decreases as the relative proliferation rate r of infected hepatocytes decreases.

There is a small range of values of r for which (3.14) has no solutions. For this region, $\mathbf{u}^*$ does not exist, and our approximation to t_d fails despite the presence of a positive but finite delay. The perturbation approximation to t_d continues to work for some of this region but also eventually fails for r just below the exact bistability threshold. Approximation of t_d in this region can be performed using dynamical systems theory, as described in Appendix C.

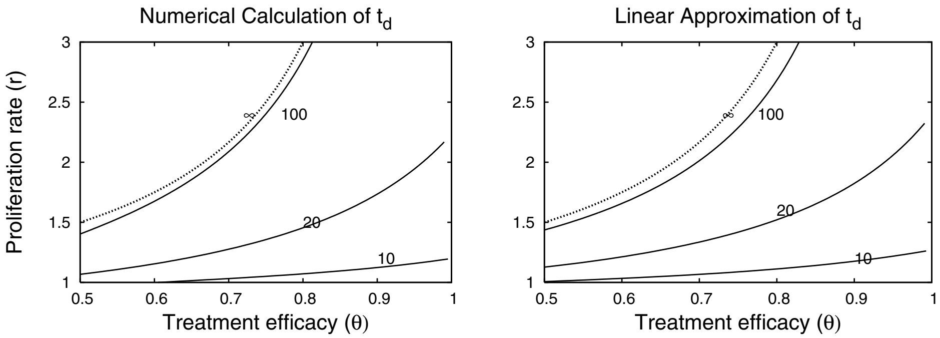


FIG. 13. *Side-by-side comparison of contour plots of the treatment delay t_d using numerical solution of system* (2.7) *(left) and the formula* (4.8) *derived from the linear approximation (right). The approximate bound on bistability, $r = d + \frac{d}{(1-\theta)b}$, labeled ∞, is the same in both plots. Contour heights are* 10*,* 20*,* 100*, and* ∞*. Parameter values $d = 0.5, b = 1, s = 10^{-3}, q = 0$. $\theta_c = 0.5$ when $r = 0$ in both plots.*

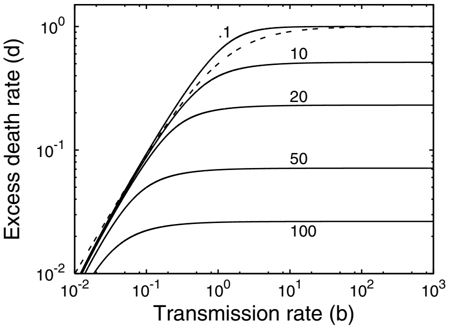


FIG. 14. *Contour plot of the treatment delay t_d from* (4.8) *as a function of d and b when $r = 1, \theta = 1, s = 10^{-3}$, and $q = 0$. Contours at* .1, 10, 20, 50, *and* 100*. The dotted line $d = rb/(1 + b)$ is an upper bound on the region of strong-delay effect.*

5. Discussion. Only about 20%–30% of HCV-infected individuals spontaneously clear the virus during the early phase of infection [44]. According to our model, when an individual is initially exposed to a small amount of virus, infection cannot be established unless the disease-free reproductive number is greater than 1. If the reproductive number is greater than 1, virus will spread among hepatocytes, eventually infecting some or all of the cells it targets. In addition, viral dynamics during this phase may be monotone or oscillatory but are expected to converge to a stationary equilibrium. Homeostatic proliferation of infected cells has only a small effect on the reproductive number, but diminishes oscillations [4] and increases the proportion of target hepatocytes infected at steady state. However, the dynamics may be bistable if the proliferation rate of infected hepatocytes is faster than that of uninfected hepatocytes. The HCV kinetics during primary infection, before the adaptive immune response against HCV is induced, both in humans [18] and chimpanzees [29], has been observed to be monotone; i.e., after a fast viral increase the virus stabilizes without observed oscillations in a high viral load steady state. This lack of observed oscillations supports our hypothesis that homeostatic proliferation of infected cells exists.

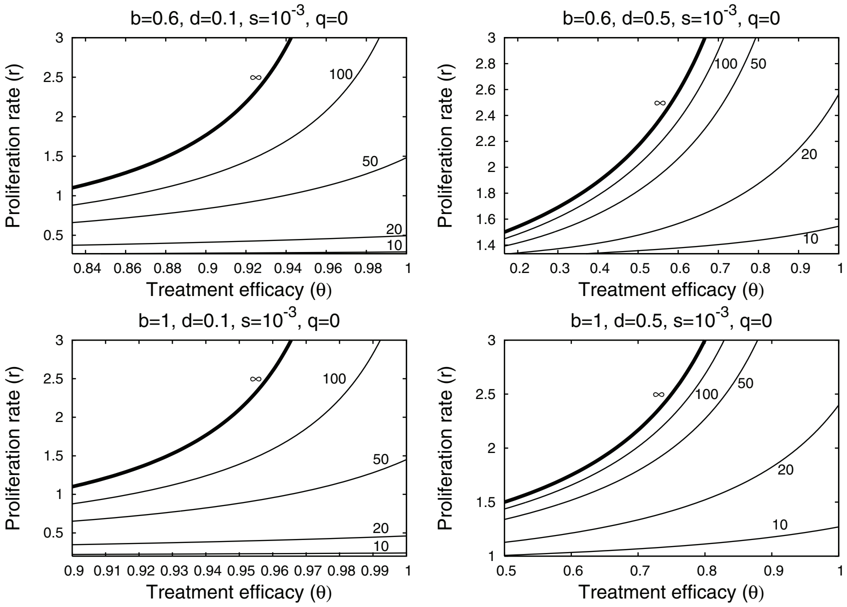


Fig. 15. *Four contour plots of treatment delay t_d in the strong-delay region of Figure* 11*, calculated from* (4.8) *when hepatocyte immigration is slow. In the parameter region above the ∞-contour, treatment is not sufficiently effective to overcome the local stability of the infected-cell population. Below the ∞-contour, treatment successfully clears infection, with the length of the delay given by the contour values. Parameter values are given at the top of each plot. The left-hand boundary in each plot corresponds to $\theta = \theta_c$ when $r = 0$.*

The typical HCV RNA decay observed during therapy with standard or pegylated interferon-α alone or in combination with ribavirin is biphasic—characterized by an initial rapid viral decline (first phase) followed by a slower decay (second phase) [37]. In about 30%–40% of treated patients triphasic viral declines have also been observed [19, 48, 1, 23]. In some patients (nonresponders) viral loads may not decline. In others, viral load initially declines (first phase) followed by maintenance of a steady level lower than baseline (flat partial responders). Here we have mathematically characterized a model of HCV dynamics [6] that encompasses the observed viral kinetic profiles under therapy. We speculate that in nonresponders the drug effectiveness, θ, may not exceed θ_p (Figure 9), and therefore viral load does not decline under therapy. Flat partial responders may be explained as a consequence of drug efficacy higher than θ_p but lower than the critical drug efficacy θ_c (Figure 9). Viral clearance occurs when $\theta > \theta_c$ via biphasic or triphasic viral decline when the hepatocyte proliferation rate, r, is lower or higher than the hepatocyte death rate, d, respectively (system (2.7) without "cure"; Figure 9).

Using perturbation theory, we showed that a delay can occur between the start of treatment and the first measurable decline in the number of infected hepatocytes under efficient therapy ($\theta > \theta_c$) because of the influence of a nearby saddle-node bifurcation in the system (Figure 12). Equation (4.8) can be used to approximate the duration of this delay. In terms of viral dynamics, this delay appears as a shoulder phase separating the initial decay in viral load at the start of treatment from the asymptotic clearance phase. One of the conditions for the existence of this delay between the initial decrease and asymptotic clearance is that the number of infected

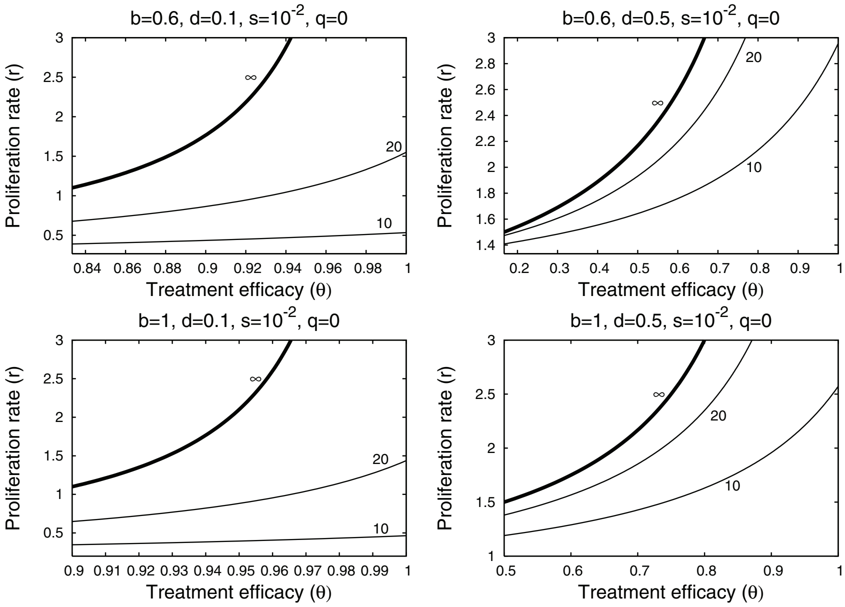


FIG. 16. *Four contour plots of treatment delay t_d in the strong-delay region of Figure* 11*, calculated from formula* (4.8)*. Parameter values are stated at the top of each plot. In the parameter region above the ∞-contour, treatment is not sufficiently effective to overcome the local stability of the infected-cell population. Below the ∞-contour, treatment successfully clears infection, with the length of the delay given by the contour values. In these plots, immigration is fast ($s = 10^{-2}$) and significantly reduces the delay before treatment reduces the number of infected cells, compared to Figure* 15*. The left-hand boundary in each plot corresponds to $\theta = \theta_c$ when $r = 0$.*

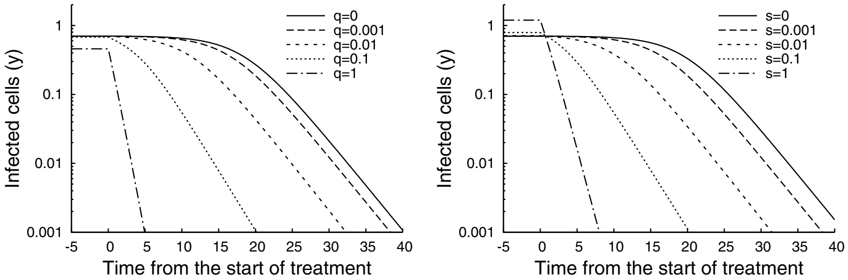


FIG. 17. *Time series plot of changes in the number of infected cells under treatment with $s = 0.001$ (left) and $q = 0.001$ (right). As the curing of infected cells q is increased from* 0 *to* 1*, the treatment delay decreases from* 15 *to* 0 *(left). Similarly, treatment delay decreases as the immigration rate s increases (right). Note that only very large values of s and q significantly affect the pretreatment state. Parameter values $d = 0.3, b = 3, \theta = 1, r = 1$.*

cells is much larger than the number of uninfected cells at the start of therapy. During therapy the number of uninfected cells increases. Because of density-dependent homeostatic processes, the proliferation of infected cells slows as the number of uninfected cells increases. When this proliferation slows to the point at which it no longer keeps up with the rate of infected cell loss, the number of infected cells start to decline. The shoulder persists until the ratio between uninfected cells and infected cells is approximately one. We found that the stopping condition, $T/I \approx 1$, for calculating when the shoulder phase ends, is simple and robust for calculating t_d over a large-shoulder parameter range. Other topping conditions $T/I \approx 1$ are discussed in Appendix B.

When using t_d in the context of the full model (system (2.1)), i.e., calculating the viral shoulder phase and hence a triphasic viral decay, our formula (4.8) has to be adjusted. Since interferon-α mainly inhibits viral production, and we assume that initially the infected cell number remains close to its level before therapy, then the model (system (2.1)) predicts that viral load will decline from its baseline value, V_0, according to the equation $V(t) = V_0\left(1 - \epsilon + \epsilon e^{-ct}\right)$ [37]. This equation for the first phase of viral decline predicts that at times long compared with $1/c$, the average free virion lifetime in serum, the viral load will decline to $(1-\epsilon)V_0$ over an interval of length $\ln(1-\epsilon)/c$ (Figure 2). Therefore, our formula (4.8), which estimates the length of time since the start of therapy and the beginning of the third phase of decline (Figure 2(right)) can be adjusted to the actual viral shoulder duration by subtracting the relaxation time $\ln(1-\epsilon)/c$ from the dimensional form of t_d.

We have previously predicted, using system (2.1), that the spontaneous curing (q) of infected cells by a noncytolytic immune response is necessary to prevent a significant loss of liver cells during acute HCV infection in chimpanzees [5]. Direct evidence for noncytolytic clearance of HCV from infected cells has not yet been found, but interferon-α has cured replicon cells [2], and clearance of hepatitis-B-virus-infected hepatocytes has been shown to occur through noncytolytic mechanisms [16]. In the context of treatment in chronic-HCV patients, our theory predicts that any shoulder phase will be shortened by a strong noncytolytic response.

HCV is the only known RNA virus with an exclusively cytoplasmic life cycle that is associated with cancer [47]. The mechanisms by which it causes cancer are unclear. It may be that the path to hepatocellular carcinoma in chronic hepatitis C shares some important features with human papillomavirus-induced carcinogenesis [17]. Interactions of HCV proteins with key regulators of the cell cycle, e.g., the retinoblastoma protein [34] and p53 [22], may lead to enhanced cellular proliferation over uninfected cells and may also compromise multiple cell cycle checkpoints that act to maintain genomic integrity [11], thus setting the stage for carcinogenesis. In light of these speculations, the proliferation of HCV-infected cells, r_I, may be higher than proliferation of uninfected cells, r_T. Therefore, in this study we also analyzed this assumption (i.e., $r_I > r_T$). We found that when $r_I > r_T$, bistability arises (under certain parameter values), and imposing $r_I < r_T$ is sufficient to preclude bistability. However, more experimental work is needed to test the consistency of this view of homeostatic proliferation with the behavior of hepatocytes in vivo. Our model makes some predictions regarding changes in the total hepatocyte numbers over the course of infection and treatment. Since liver function is correlated with hepatocyte numbers, the total number of hepatocytes may be an important medical indicator and may further inform our understanding of HCV.

On a mathematical note, there is as yet no global stability analysis of system (2.1). Of particular importance, a closer analysis of the quasi-steady-state approximation is needed. This is emphasized by numerical observations that a Hopf bifurcation of the partial-infection stationary solution can occur if the viral clearance rate c is not sufficiently large. Applications and extension of methods from De Leenheer and Smith [9], De Leenheer and Pilyugin [8], and Korobeinikov [24] may prove useful in further work.

The analyses presented here not only are important for HCV infection but also should be relevant for modeling other infections with hepatotropic viruses, such as hepatitis B virus. Many mathematical models for the study of hepatitis B virus DNA kinetics under therapy ignore the proliferation of virus-infected cells [36]. Interestingly, besides the typical biphasic decay in viral load, other kinetic profiles have been observed, such as triphasic. As our model allows one to predict more complex viral decay profiles, we hope that it will be useful for understanding complex HCV and HBV kinetics under therapy [7].

Appendix A. Perturbation approximations. Regular perturbation methods can be used to approximate the stationary solutions of system (2.7) in the limits of small s and q, based on the polynomials in (3.11) and (3.14a). Equation (3.11) is independent of q, so let $x = x_0 + sx_s + o(s)$. Substituting into (3.11), $(x_0 + sx_s)^2 - (x_0 + sx_s) = s$. Collecting like terms,

$$x_0(x_0 - 1) - (1 - 2x_0x_s + x_s)s + o(s) = 0. \tag{A.1}$$

From the zeroth-order term in s, $x_0 \in \{0, 1\}$, and to first order, $x_s = \frac{1}{2x_0-1}$. Thus, the two corresponding stationary solutions for small s and q are

$$(-s + o(s), 0) \quad \text{and} \quad (1 + s + o(s), 0). \tag{A.2}$$

Equation (3.14a) depends on both s and q, so let $x = x_0 + sx_s + qx_q + o(s, q)$. Substituting into (3.14a) and collecting like terms,

$$\begin{aligned} 0 = {} & b(1 + b - r)x_0^2 + (rb - db - d)x_0 \\ & - (-x_srb + x_sdb + x_sd - 2bx_0x_s - 2b^2x_0x_s + 2bx_0x_sr + r)s \\ & - (-bx_qr + x_qdb + x_qd - x_0r + x_0 + 2x_0b - 2(1 + b + r)bx_0x_q + r - d)q. \end{aligned} \tag{A.3}$$

To highest order,

$$x_0 \in \left\{0, \frac{d + db - rb}{b(1 + b - r)}\right\}. \tag{A.4}$$

The first-order corrections in s and q are

$$x_s = \frac{r}{2x_0b + 2b^2x_0 - 2bx_0r - d - bd + br}, \tag{A.5a}$$

$$x_q = \frac{r - x_0r + x_0 + 2x_0b - d}{2x_0b + 2b^2x_0 - 2bx_0r - d - bd + br}. \tag{A.5b}$$

Equation (3.14b) can then be used to determine y. In the special case of $x_0 = 0$, the stationary solution is given by (3.15). This can be used to approximate both the pretreatment and posttreatment (substituting $(1 - \theta)b$ for b) stationary solutions when applying (4.8).

Appendix B. Choosing a treatment delay threshold. Using (4.4), we can approximate the delay, t_d, before the number of infected hepatocytes begins to decline. But to do this, we have to find a quantitative rule for determining the end of the treatment delay. One way to do this is to choose a line, represented by a vector $\mathbf{k}$ and a constant k_0, such that the shoulder ends when the approximate solution intersects this line. Thus, t_d is defined such that

$$\mathbf{k}^T\mathbf{u}(t_d) = k_0. \tag{B.1}$$

Table B.1
Possible stopping condition choices for calculation of t_d.

Description	$\mathbf{k}$	k_0	Comment
Upper bound	$[1,1]$	1	$\mathbf{u}(t)$ may not intersect
$x(t) = y(t)$	$[1,-1]$	0	$\mathbf{u}(t)$ may not intersect
90% threshold	$[0,1]$	$.9\,(1-d/r)$	$\mathbf{u}(t)$ may not intersect
Uninfected cells only	$[1,0]$	$\dfrac{d}{r}$	$t_d \to 0$ as $d \to 0$, although the delay may not
Uninfected cells only	$[1,0]$	$.1$	May not correspond to the full delay

Since we are concerned only with the divergence from steady state, we can ignore the stable mode of (4.4), and then (B.1) leads to the formula

(B.2)
$$t_d = \frac{r}{d + (1-\theta)b(d-r)} \log\left\{ \frac{[(r-(1-\theta)b)(r-d)+d](\mathbf{k}_1\mathbf{u}_1^* + \mathbf{k}_2\mathbf{u}_2^* - k_0)}{[(\mathbf{k}_1 - \mathbf{k}_2)(r-(1-\theta)b)(r-d) + \mathbf{k}_1 d](\mathbf{u}_1^* - \mathbf{u}_1(0))} \right\}.$$

Several choices for $\mathbf{k}$ and k_0 are summarized in Table B.1, along with their drawbacks. The unstable manifold of $\mathbf{u}^*$ has the initial direction

$$\big[d - (r-d)((1-\theta)b - r), \quad (r-d)((1-\theta)b - r)\big]. \tag{B.3}$$

If we constrain the application of (B.2) to the region of Figure 11 where $\theta > \theta_c$ and $r > d$, we can show that the choice of $\mathbf{k} = [1,-1]$, $k_0 = 0$ always gives a solution for t_d. This is because the first component of the eigenvector is positive and the second is negative, ensuring that the orbit approximated by a line in the direction of the eigenvector will always intersect the line $y = x$. Numerical evidence indicates that this choice is reasonably consistent with the qualitative character of delays, and we will use it throughout this paper. However, it underestimates the delay time in cases where $r - d$ is small. In such situations, $\mathbf{k} = [1,1]$, $k_0 = 1$ gives better approximations to t_d.

Appendix C. Bottle neck calculations. When the growth rate r is slightly smaller than the critical value r^* that introduces bistability (Figure 11), (4.4) cannot be used to estimate the treatment delay t_d because there is no nearby equilibrium around which we can linearize. However, we can still approximate the treatment delay by transforming the system near the bifurcation point into normal form [15]. The normal form of a generic saddle-node bifurcation satisfies the first-order differential equation

$$\dot{u} = a_0(r^* - r) + a_2 u^2, \tag{C.1}$$

where r is the bifurcation parameter and $r = r^*$ is the bifurcation point with $a_0 > 0$ and $a_2 \neq 0$. Using elementary integration methods, we can show that for $r \approx r^*$ the time it takes for a solution to pass from a negative initial position to a positive final position, both far from the origin, is approximately given by

$$t_d \approx \sqrt{\pi^2/[a_0 a_2 (r^* - r)]}. \tag{C.2}$$

As r is increased toward r^*, the time becomes longer. If $r > r^*$, the time is infinite because solutions are trapped by an intermediate attracting state. For this HCV model, our task to calculate r^*, a_0, and a_2 by transforming system (2.7) into normal-form near the saddle-node bifurcation that introduces bistability.

We can determine the bifurcation point r^* by setting the discriminant of x in (3.14a) (with transmission rate $(1-\theta)b$ instead of b) equal to zero. The result is a quadratic polynomial for r^*, where the smaller solution corresponds to a saddle-node bifurcation for nonbiological values of x, and the larger solution corresponds to bifurcation which is biologically important.

Once r^* is known, we calculate the feasible nonhyperbolic equilibrium solution $(x^*(r^*), y^*(r^*))$ using (3.14) and transform system (2.7) using a change of variables of the form

$$x = x^* + M_{xu}u + M_{xv}v + M_{xr}(r - r^*), \tag{C.3a}$$

$$y = y^* + M_{yu}u + M_{yv}v + M_{yr}(r - r^*) \tag{C.3b}$$

such that locally the system has the form

$$\dot{u} = a_0(r^* - r) + a_2u^2 + o(r^* - r, u^2) + O(v), \tag{C.4a}$$

$$\dot{v} = -a_3v + o(v), \tag{C.4b}$$

where a_0 and a_2 satisfy the conditions given above and $a_3 > 0$. This transformation can be performed using the eigenvalue decomposition of the Jacobian at the equilibrium point and then choosing M_{xr} and M_{yr} to eliminate extra terms in $\dot{u}$ and $\dot{v}$. The $O(v)$ terms are neglected because v converges to 0 exponentially near the bifurcation. The procedure is easily implemented numerically, but we have not produced a simple analytic formula for the result.

Acknowledgments. T. Reluga thanks A. Zilman for helpful discussion concerning the calculations in Appendix C.

REFERENCES

[1] F. C. Bekkering, A. U. Neumann, J. T. Brouwer, R. S. Levi-Drummer, and S. W. Schalm, *Changes in anti-viral effectiveness of interferon after dose reduction in chronic hepatitis C patients: A case control study*, BMC Gastroenterology, 1 (2001), p. 14.

[2] K. J. Blight, J. A. McKeating, and C. M. Rice, *Highly permissive cell lines for subgenomic and genomic hepatitis C virus RNA replication*, J. Virology, 76 (2002), pp. 13001–13014.

[3] H. Dahari, A. Feliu, M. Garcia-Retortillo, X. Forns, and A. U. Neumann, *Second hepatitis C replication compartment indicated by viral dynamics during liver transplantation*, J. Hepatology, 42 (2005), pp. 491–498.

[4] H. Dahari, A. Lo, R. M. Ribeiro, and A. S. Perelson, *Modeling hepatitis C virus dynamics: Liver regeneration and critical drug efficacy*, J. Theoret. Biol., 47 (2007), pp. 371–381.

[5] H. Dahari, M. Major, X. Zhang, K. Mihalik, C. M. Rice, A. S. Perelson, S. M. Feinstone, and A. U. Neumann, *Mathematical modeling of primary hepatitis C infection: Noncytolytic clearance and early blockage of virion production*, Gastroenterology, 128 (2005), pp. 1056–1066.

[6] H. Dahari, R. M. Ribeiro, and A. S. Perelson, *Triphasic decline of HCV RNA during antiviral therapy*, Hepatology, 46 (2007), pp. 16–21.

[7] H. Dahari, E. Shudo, R. M. Ribeiro, and A. S. Perelson, *Modeling complex decay profiles of hepatitis B virus during antiviral therapy*, Hepatology, to appear (DOI: 10.1002/hep.22586).

[8] P. De Leenheer and S. Pilyugin, *Multi-strain Virus Dynamics with Mutations: A Global Analysis*, preprint, arXiv:0707.4501.

[9] P. De Leenheer and H. Smith, *Virus dynamics: A global analysis*, SIAM J. Appl. Math., 63 (2003), pp. 1313–1327.

[10] G. Di Liberto and C. Féray, *The anhepatic phase of liver transplantation as a model for measuring the extra-hepatic replication of hepatitis C virus*, J. Hepatology, 42 (2005), pp. 441–443.

[11] S. Duensing and K. Munger, *Mechanisms of genomic instability in human cancer: Insights from studies with human papillomavirus oncoproteins*, Internat. J. Cancer, 109 (2004), pp. 157–162.

[12] N. Fausto, *Liver regeneration and repair: Hepatocytes, progenitor cells, and stem cells*, Hepatology, 39 (2004), pp. 1477–1487.

[13] M. W. Fried, M. L. Shiffman, K. R. Reddy, C. Smith, G. Marinos, F. L. Goncales, D. Haussinger, M. Diago, G. Carosi, D. Dhumeaux, A. Carxi, A. Lin, J. Hoffman, and J. Yu, *Peginterferon alfa-*2*a plus ribavirin for chronic hepatitis C virus infection*, New England Journal of Medicine, 347 (2002), pp. 975–982.

[14] H. Gómez-Acevedo and M. Y. Li, *Backward bifurcation in a model for HTLV-*I *infection of CD*4+ *T cells*, Bull. Math. Biol., 67 (2005), pp. 101–114.

[15] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[16] L. G. Guidotti, R. Rochford, J. Chung, M. Shapiro, R. Purcell, and F. V. Chisari, *Viral clearance without destruction of infected cells during acute HBV infection*, Science, 284 (1999), pp. 825–829.

[17] C. M. Hebner and L. A. Laimins, *Human papillomaviruses: Basic mechanisms of pathogenesis and oncogenicity*, Rev. Med. Virology, 16 (2006), pp. 83–97.

[18] B. L. Herring, R. Tsui, L. Peddada, M. Busch, and E. L. Delwart, *Wide range of quasispecies diversity during primary hepatitis C virus infection*, J. Virology, 79 (2005), pp. 4340–4346.

[19] E. Herrmann, J. Lee, G. Marinos, M. Modi, and S. Zeuzem, *Effect of ribavirin on hepatitis C viral kinetics in patients treated with pegylated interferon*, Hepatology, 37 (2003), pp. 1351–1358.

[20] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[21] G. E. Hutchinson, *An Introduction to Population Ecology*, Yale University Press, New Haven, CT, 1978.

[22] C. F. Kao, S. Y. Chen, J. Y. Chen, and Y. H. W. Lee, *Modulation of* p53 *transcription regulatory activity and post-translational modification by hepatitis C virus core protein*, Oncogene, 23 (2004), pp. 2472–2483.

[23] T. L. Kieffer, C. Sarrazin, J. S. Miller, M. W. Welker, N. Forestier, A. D. Kwong, and S. Zeuzem, *Telaprevir and pegylated interferon-alpha-*2*a inhibit wild-type and resistant genotype* 1 *hepatitis C virus replication in patients*, Hepatology, 46 (2007), pp. 631–639.

[24] A. Korobeinikov, *Global properties of basic virus dynamics models*, Bull. Math. Biol., 66 (2004), pp. 879–883.

[25] B. D. Lindenbach and C. M. Rice, *Unravelling hepatitis C virus replication from genome to function*, Nature, 436 (2005), pp. 933–938.

[26] R. A. Macdonald, *"Lifespan" of liver cells. Autoradio-graphic study using tritiated thymidine in normal, cirrhotic, and partially hepatectomized rats*, Arch. Internal Medicine, 107 (1961), pp. 335–343.

[27] I. R. Mackay, *Hepatoimmunology: A perspective. Special Feature*, Immunol. Cell Biol., 80 (2002), pp. 36–44.

[28] R. N. M. Macsween, A. D. Burt, B. C. Portmann, K. G. Ishak, P. J. Scheuer, and P. P. Anthony, *Pathology of the Liver*, Churchill Livingstone, London, 1987.

[29] M. E. Major, H. Dahari, K. Mihalik, M. Puig, C. M. Rice, A. U. Neumann, and S. M. Feinstone, *Hepatitis C virus kinetics and host responses associated with disease and outcome of infection in chimpanzees*, Hepatology, 39 (2004), pp. 1709–1720.

[30] M. P. Manns, J. G. McHutchison, S. C. Gordon, V. K. Rustgi, M. Shiffman, R. Reindollar, Z. D. Goodman, K. Koury, M. H. Ling, and J. K. Albrecht, *Peginterferon alfa-*2*b+ ribavirin compared with interferon alfa-*2*b+ ribavirin for initial treatment of chronic hepatitis C: A randomised trial*, The Lancet, 358 (2001), pp. 958–965.

[31] P. Meuleman, L. Libbrecht, R. De Vos, B. de Hemptinne, K. Gevaert, J. Vandekerckhove, T. Roskams, and G. Leroux-Roels, *Morphological and biochemical characterization of a human liver in a uPA-SCID mouse chimera*, Hepatology, 41 (2005), pp. 847–856.

[32] G. K. Michalopoulos and M. C. DeFrances, *Liver regeneration*, Science, 276 (1997), pp. 60–66.

[33] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, Twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.

[34] T. Munakata, M. Nakamura, Y. Liang, K. Li, and S. M. Lemon, *Down-regulation of the retinoblastoma tumor suppressor by the hepatitis C virus NS*5*B RNA-dependent RNA polymerase*, Proc. Nat. Acad. Sci. USA, 102 (2005), pp. 18159–18164.

[35] J. D. Murray, *Mathematical Biology*, 2nd ed., Springer, New York, 1993.

[36] A. U. Neumann, *Hepatitis B viral kinetics: A dynamic puzzle still to be resolved*, Hepatology, 42 (2005), pp. 249–254.

[37] A. U. Neumann, N. P. Lam, H. Dahari, D. R. Gretch, T. E. Wiley, T. J. Layden, and A. S. Perelson, *Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy*, Science, 282 (1998), pp. 103–107.

[38] NIH, *National Institutes of Health Consensus Development Conference: Management of hepatitis C*, Hepatology, 36 (2002), pp. S3–S20.

[39] I. M. Pedersen, G. Cheng, S. Wieland, S. Volinia, C. M. Croce, F. V. Chisari, and M. David, *Interferon modulation of cellular microRNAs as an antiviral mechanism*, Nature, 449 (2007), pp. 919–922.

[40] A. Perelson, A. Neumann, M. Markowitz, J. Leonard, and D. Ho, *HIV*-1 *dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time*, Science, 271 (1996), pp. 1582–1586.

[41] A. S. Perelson, *Modelling viral and immune system dynamics*, Nature Rev. Immunol., 2 (2002), pp. 28–36.

[42] A. S. Perelson, E. Herrmann, F. Micol, and S. Zeuzem, *New kinetic models for the hepatitis C virus*, Hepatology, 42 (2005), pp. 749–754.

[43] K. A. Powers, R. M. Ribeiro, K. Patel, S. Pianko, L. Nyberg, P. Pockros, A. J. Conrad, J. McHutchison, and A. S. Perelson, *Kinetics of hepatitis C virus reinfection after liver transplantation*, Liver Transplantation, 12 (2006), pp. 207–216.

[44] B. Rehermann and M. Nascimbeni, *Immunology of hepatitis B virus and hepatitis C virus infection*, Nature Rev. Immunol., 5 (2005), pp. 215–229.

[45] R. Ribeiro, *Dynamics of alanine aminotransferase during hepatitis C virus treatment*, Hepatology, 38 (2003), pp. 509–517.

[46] L. B. Seeff, *Natural history of chronic hepatitis C*, Hepatology, 36 (2002), pp. S35–S46.

[47] L. B. Seeff and J. H. Hoofnagle, *Epidemiology of hepatocellular carcinoma in areas of low hepatitis B and hepatitis C endemicity*, Oncogene, 25 (2006), pp. 3771–3777.

[48] R. E. Sentjens, C. J. Weegink, M. G. Beld, M. C. Cooreman, and H. W. Reesink, *Viral kinetics of hepatitis C virus RNA in patients with chronic hepatitis C treated with* 18 *MU of interferon alpha daily*, European J. Gastroenterology and Hepatology, 14 (2002), pp. 833–840.

[49] S. Sherlock and J. Dooley, *Diseases of the Liver and Biliary System*, Blackwell Science, Boston, 2002.

[50] N. D. Theise, M. Nimmakayalu, R. Gardner, P. B. Illei, G. Morgan, L. Teperman, O. Henegariu, and D. S. Krause, *Liver from bone marrow in humans*, Hepatology, 32 (2000), pp. 11–16.

[51] R. Thimme, D. Oldach, K. M. Chang, C. Steiger, S. C. Ray, and F. V. Chisari, *Determinants of viral clearance and persistence during acute hepatitis C virus infection*, J. Exper. Med., 194 (2001), pp. 1395–1406.

[52] X. Wei, S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw, *Viral dynamics in human immunodeficiency virus type* 1 *infection*, Nature, 373 (1995), pp. 117–122.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 1024–1042

THE IDENTIFICATION OF THIN DIELECTRIC OBJECTS FROM FAR FIELD OR NEAR FIELD SCATTERING DATA*

NOAM ZEEV† and FIORALBA CAKONI†

Abstract. We consider the inverse scattering problem of determining the shape and the material properties of a thin dielectric infinite cylinder having an open arc as cross section from knowledge of the TM-polarized scattered electromagnetic field at a fixed frequency. We investigate two reconstruction approaches, namely the linear sampling method and the reciprocity gap functional method, using far field or near field data, respectively. Numerical examples are given showing the efficaciousness of our algorithms.

Key words. direct and inverse scattering, scattering from cracks, linear sampling method, electromagnetic scattering, reciprocity gap functional method, thin dielectric objects

AMS subject classifications. 35R30, 35Q60, 35J40, 78A25

DOI. 10.1137/070711542

1. Introduction.

Important problems in nondestructive evaluation include the detection of flaws in materials in specific (typically thin) areas, as well as the determination of the integrity of thin coatings. We refer the reader to the special issue of Inverse Problems [22] for a detailed account on different approaches to these problems using electromagnetic waves. In particular, considerable work has been done on inversion schemes using eddy-current approximation measurements to detect the presence of thin anomalies [3], [13], [29]. In related work [4], [24], the authors use electrostatic and electromagnetic measurements, respectively, to detect the shape of a thin target. In addition to the shape, it is of course desirable to obtain information on the material properties of the target. In this paper we show the applicability of qualitative methods in inverse scattering [7] to these problems. In particular, we investigate the inverse problem of using far field or near field time harmonic electromagnetic measurements to determine the shape and information about the thickness and physical properties of a thin dielectric film embedded in a known inhomogeneous background. Such problems arise in the study of optical devices in communication networks [27] (typical structures of this type can be found in [14]) or in the detection of thin air pockets inside structures [29]. In this work we assume that the obstacle is a thin dielectric right cylinder whose properties depend only on the cross section of the cylinder and that the incident electromagnetic field is E-polarized. After factoring out the term $e^{-i\omega t}$, where ω is the fixed frequency, the only nonzero component u of the total electric field satisfies the Helmholtz equation

$$\Delta u + k^2 n(x) u = 0$$

in the exterior of the cylinder, where the complex valued function $n(x)$ is the index of refraction of the background medium which satisfies $\operatorname{Re} n > 0$ and $\operatorname{Im} n \geq 0$. Difficulties arise in computing the total field inside the thin dielectric obstacle due to

[*] Received by the editors December 20, 2007; accepted for publication (in revised form) September 12, 2008; published electronically January 14, 2009.
http://www.siam.org/journals/siap/69-4/71154.html

[†] Department of Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (zeev@math.udel.edu, cakoni@math.udel.edu). The research of the second author was supported in part by the U.S. Air Force Office of Scientific Research under grant FA9550-05-1-0127.

the fact that the index of refraction and the thickness of the obstacle are of different scales. The direct scattering problem for a thin dielectric structure was studied in [1] and [27] where a perturbation approach was used to approximate the solution by solving a sequence of integral equations. Alternatively, based on asymptotic analysis with respect to the thickness of the obstacle (see [16]), a first approximate model of the wave propagation inside the obstacle is to replace the obstacle by an infinite cylinder having an open arc in $\mathbb{R}^2$ as its cross section and the interior field by an appropriate boundary condition on the arc. Both the perturbation method and the arc approximation model are based on asymptotic analysis of the exact model with respect to the thickness, and they therefore compute an approximation to the total field. The error analysis of the forward problem for the perturbation technique can be found in [1] and [27], and for the arc approximation model in [16].

Our analysis of the inverse problem uses the arc approximation model. Note that the physical properties and the thickness of the thin dielectric obstacle appear now as a boundary coefficient. We remark that this model is well suited to our inversion algorithm, especially since the noniterative inversion methods such as the linear sampling method and the reciprocity gap functional method are able to reconstruct boundary coefficients in addition to the support. More specifically, let h be the thickness and Γ the cross section of the mean surface of the dielectric medium. Assuming that the interior magnetic field is approximated up to $O(h)$ error, whereas the interior electric field is approximated up to $O(h^2)$ error, the following boundary conditions on the open arc Γ are obtained for the component u of the total electric field [15]:

$$\left[\frac{\partial u}{\partial \nu}\right] = 0 \qquad \text{and} \qquad [u] - i\lambda \frac{\partial u^+}{\partial \nu} = 0 \qquad \text{on } \Gamma,$$

where $u^{\pm}(x) = \lim_{h\to 0^+} u(x \pm h\nu)$ and $\frac{\partial u^{\pm}}{\partial \nu}(x) = \lim_{h\to 0^+} \nu \cdot \nabla u(x \pm h\nu)$ for $x \in \Gamma$, $[u] := u^+ - u^-$ and $\left[\frac{\partial u}{\partial \nu}\right] := \frac{\partial u^+}{\partial \nu} - \frac{\partial u^-}{\partial \nu}$ are the respective jumps across Γ, and the dimensionless positive valued function $\lambda > \lambda_0 > 0$ involves electric permittivity and magnetic permeability of the dielectric medium and the background as well as the thickness h and frequency ω. Note that the above condition can fail at the tips of the crack. Here we assume that $\Gamma \subset \mathbb{R}^2$ is a *simple piecewise smooth arc*, i.e., $\Gamma = \{\rho(s) : s \in [s_0, s_1]\}$, where the mapping $\rho : [s_0, s_1] \to \mathbb{R}^2$ is one-to-one, continuous, and piecewise smooth. The normal vector ν pointing to the right side of Γ is defined everywhere except at a finite number of points on Γ.

Hence we arrive at the following boundary value problem for the scattered field u^s due to an incident field u^i scattered by the crack Γ:

$$\Delta u^s + k^2 n(x) u^s = 0 \qquad \text{in } \mathbb{R}^2 \setminus \overline{\Gamma}, \tag{1.1}$$

$$\left[\frac{\partial (u^s + u^i)}{\partial \nu}\right] = 0 \qquad \text{on } \Gamma, \tag{1.2}$$

$$\left[(u^s + u^i)\right] - i\lambda \frac{\partial (u^s + u^i)^+}{\partial \nu} = 0 \qquad \text{on } \Gamma, \tag{1.3}$$

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^s}{\partial r} - iku^s\right) = 0, \tag{1.4}$$

where the Sommerfeld radiation condition (1.4) is satisfied uniformly in $\hat{x} = x/|x|$ with $r = |x|$. Here we assume that, in general, the positive index of refraction $n(x) > 0$ for the background medium satisfies $n(x) = 1$ outside a large ball containing the crack and k is the wave number in the air. In this study the incident field can be a plane wave or the field generated by a point source, and this will become precise later.

The main concern of this paper is to solve the inverse problem of determining the shape Γ and some information on λ from measured far field or near field scattered data. In order to develop the mathematical tools to study the inverse problem, in the next section we investigate the well-posedness of the direct scattering problem (1.1)–(1.4). We apply a boundary integral equation method to obtain a Fredholm first kind integral equation on Γ for the scattered field. In section 3 we formulate and solve the inverse scattering problem using far field scattering data due to plane waves as incident fields. For simplicity in this section we assume that the crack is embedded in a homogeneous background; i.e., $n = 1$ everywhere. We apply the *linear sampling method*, which was first introduced in [6] for the case of an obstacle with empty interior, to determine the shape of the crack. After the reconstruction of Γ (without making a priori use of the boundary condition) we use the solution of the far field equation to reconstruct λ as well. For information on other solution methods for the inverse scattering problem for Dirichlet or Neumann cracks from far field data, we refer the reader to [2], [17], [18], and [21]. In section 4 we consider the case when the crack is embedded in a known inhomogeneous background and the data is the scattered field measured on a closed curve surrounding the crack due to a point source as incident field. We modify the reciprocity gap functional method which up to now has been developed only for obstacles with nonempty interior [8], [12]. The last section of this paper is dedicated to numerical implementation with examples of both algorithms for solving the inverse problem. We note that the solution of the inverse problem in both cases is based on solving an ill-posed linear equation whose right-hand side involves normal derivatives with respect to the unknown crack. We propose a new approach to deal with this difficulty which was left as an open question in [6].

2. The solution of the direct scattering problem.

In order to formulate the above scattering problems more precisely we need to properly define the trace spaces on Γ. To this end we extend the arc Γ to an arbitrary piecewise smooth, simply connected, closed curve ∂D enclosing a bounded domain D such that the normal vector ν on Γ coincides with the outward normal vector on ∂D, which we again denote by ν. The classical reference for the trace spaces is [23], and the notation there is different from those in [25]. However, in this work we use the notation in [25], because this is our main reference for the potential theory needed here. If $H^1_{loc}(\mathbb{R}^2)$, $L^2(\partial D)$, $H^{\frac{1}{2}}(\partial D)$, and $H^{-\frac{1}{2}}(\partial D)$ denote the usual Sobolev spaces, we define the following spaces:

$$\begin{aligned} L^2(\Gamma) &:= \{u|_\Gamma : u \in L^2(\partial D)\},\\ H^{\frac{1}{2}}(\Gamma) &:= \{u|_\Gamma : u \in H^{\frac{1}{2}}(\partial D)\},\\ \tilde{H}^{\frac{1}{2}}(\Gamma) &:= \{u \in H^{\frac{1}{2}}(\Gamma) : \operatorname{supp} u \subseteq \overline{\Gamma}\}. \end{aligned}$$

In other words, $\tilde{H}^{\frac{1}{2}}(\Gamma)$ contains functions $u \in H^{\frac{1}{2}}(\Gamma)$ such that their extension by zero to the whole boundary ∂D is in $H^{\frac{1}{2}}(\partial D)$ (Theorem 3.33 in [25]). (For the reader's convenience we remark that $\tilde{H}^{\frac{1}{2}}(\Gamma)$ coincides with the space $H^{\frac{1}{2}}_{00}(\Gamma)$ introduced by Lions and Magenes (see [23, p. 66]).) Now we denote by $H^{-\frac{1}{2}}(\Gamma)$ the dual space of $\tilde{H}^{\frac{1}{2}}(\Gamma)$ and by $\tilde{H}^{-\frac{1}{2}}(\Gamma)$ the dual space of $H^{\frac{1}{2}}(\Gamma)$. Hence we have the chain

$$\mathcal{D}(\Gamma) \subset \tilde{H}^{\frac{1}{2}}(\Gamma) \subset H^{\frac{1}{2}}(\Gamma) \subset L^2(\Gamma) \subset \tilde{H}^{-\frac{1}{2}}(\Gamma) \subset H^{-\frac{1}{2}}(\Gamma) \subset \mathcal{D}'(\Gamma),$$

where $\mathcal{D}(\Gamma) := C_0^\infty(\Gamma)$. We note that $\tilde{H}^{-\frac{1}{2}}(\Gamma)$ can also be identified with $H_{\overline{\Gamma}}^{-\frac{1}{2}}(\partial D) := \{u \in H^{-\frac{1}{2}}(\partial D) :\ \operatorname{supp} u \subset \overline{\Gamma}\}$ (see Theorem 3.29 in [25]).

The scattering problem (1.1)–(1.4) is a particular case of the following boundary value problem: Let $n(x)$ be a piecewise smooth complex valued function with piecewise continuous jump discontinuities such that $\operatorname{Re} n > 0$, $\operatorname{Im} n \geq 0$, and $n(x) = 1$ outside a large enough ball, whereas λ is a piecewise smooth function on Γ such that $\lambda(x) > \lambda_0 > 0$. Given $f \in H^{-\frac{1}{2}}(\Gamma)$ and $h \in H^{-\frac{1}{2}}(\Gamma)$, find $v \in H^1_{loc}(\mathbb{R}^2 \setminus \overline{\Gamma})$ satisfying

$$\Delta v + k^2 n(x) v = 0 \qquad \text{in } \mathbb{R}^2 \setminus \overline{\Gamma}, \tag{2.1}$$

$$\left[\frac{\partial v}{\partial \nu}\right] = f \qquad \text{on } \Gamma, \tag{2.2}$$

$$[v] - i\lambda \frac{\partial v^+}{\partial \nu} = h \qquad \text{on } \Gamma, \tag{2.3}$$

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial v}{\partial r} - ikv\right) = 0. \tag{2.4}$$

THEOREM 2.1. *The problem* (2.1)–(2.4) *has at most one solution.*

Proof. Denote by B_R a sufficiently large ball with radius R containing $\overline{D}$ and by ∂B_R its boundary. Let v be a solution of (2.1)–(2.4) with $f = h = 0$. Obviously $v \in H^1(B_R \setminus \overline{D}) \cup H^1(D)$ satisfies the Helmholtz equation in $B_R \setminus \overline{D}$ and D and the following transmission conditions on the complementary part $\partial D \setminus \overline{\Gamma}$ of ∂D:

$$v^+ = v^- \qquad \text{and} \qquad \frac{\partial v^+}{\partial \nu} = \frac{\partial v^-}{\partial \nu} \qquad \text{on } \partial D \setminus \overline{\Gamma}, \tag{2.5}$$

where the + denotes the limit approaching ∂D from inside D and − the limit approaching ∂D from outside of D. An application of Green's formula for u and $\overline{u}$ in D and $B_R \setminus \overline{D}$ and using the transmission conditions (2.5) yields

$$\int_{\partial B_R} v \frac{\partial \bar{v}}{\partial \nu} dx = \int_{B_R \setminus \bar{D}} |\nabla v|^2\, dx + \int_D |\nabla v|^2\, dx - k^2 \int_{B_R\setminus\bar{D}} \overline{n}\, |v|^2\, dx$$
$$- k^2 \int_D \overline{n}\, |v|^2\, dx + \int_\Gamma [v] \frac{\partial \bar{v}}{\partial \nu} dx.$$

Using the boundary conditions (2.2)–(2.3), we now obtain

$$\int_{\partial B_R} v \frac{\partial \bar{v}}{\partial \nu} dx = \int_{B_R \setminus \bar{D}} |\nabla v|^2\, dx + \int_D |\nabla v|^2\, dx - k^2 \int_{B_R\setminus\bar{D}} \overline{n}\, |v|^2\, dx$$
$$- k^2 \int_D \overline{n}\, |v|^2\, dx + i \int_\Gamma \lambda \left|\frac{\partial v}{\partial \nu}\right|^2 dx. \tag{2.6}$$

Since $\lambda > 0$ and $\operatorname{Im} \overline{n} \leq 0$, we conclude that

$$\operatorname{Im}\left(\int_{\partial B_R} v \frac{\partial \bar{v}}{\partial \nu} dx\right) \geq 0,$$

whence from [10, Theorem 2.12] and a unique continuation argument we obtain that $v = 0$ in $\mathbb{R}^2 \setminus \overline{\Gamma}$. □

THEOREM 2.2. *The problem* (2.1)–(2.4) *has a unique solution* v *which satisfies*

$$\|v\|_{H^1(B_R\setminus\overline{\Gamma})} \leq C\left(\|f\|_{H^{-\frac{1}{2}}(\Gamma)} + \|h\|_{H^{-\frac{1}{2}}(\Gamma)}\right), \quad x \in \mathbb{R}^2 \setminus \overline{\Gamma}, \tag{2.7}$$

where the positive constant C depends on R but not on f and h.

Proof. First we note that if $v \in H^1_{loc}(\mathbb{R}^2 \setminus \overline{\Gamma})$ is a solution to (2.1)–(2.4), then $[v] \in H^{\frac{1}{2}}(\partial D)$ and $\left[\frac{\partial v}{\partial \nu}\right] \in H^{-\frac{1}{2}}(\partial D)$. Now by local regularity for solutions of the Helmholtz equation we have that $v \in C^\infty$ away from Γ, whence $[v] = \left[\frac{\partial v}{\partial \nu}\right] = 0$ on $\partial D \setminus \overline{\Gamma}$. Therefore $[v] \in \tilde{H}^{\frac{1}{2}}(\Gamma)$ and $\left[\frac{\partial v}{\partial \nu}\right] \in \tilde{H}^{-\frac{1}{2}}(\Gamma)$. Let $\mathbb{G}(x,y)$ be the radiating Green function of the background medium that satisfies

$$\Delta\mathbb{G}(x,\, y) + k^2 n(x)\mathbb{G}(x,\, y) = \delta(x-y). \tag{2.8}$$

From the Green representation formula (see [25]) we have

$$v(x) = \begin{cases} \displaystyle\int_{\partial D} \frac{\partial v^+(y)}{\partial \nu_y}\mathbb{G}(x,y)ds_y - \int_{\partial D} v^+(y)\frac{\partial}{\partial \nu_y}\mathbb{G}(x,y)ds_y, & x \in D, \\ \displaystyle -\int_{\partial D} \frac{\partial v^-(y)}{\partial \nu_y}\mathbb{G}(x,y)ds_y + \int_{\partial D} v^-(y)\frac{\partial}{\partial \nu_y}\mathbb{G}(x,y)ds_y, & x \in \mathbb{R}^2 \setminus \overline{D}, \end{cases} \tag{2.9}$$

where D is the region bounded by the extension ∂D of Γ, and the $+$ sign denotes the limit approaching ∂D from inside D whereas $-$ denotes the limit approaching ∂D from outside of D. Using the jump relations of the single- and double-layer potentials across the boundary ∂D [25], eliminating the integrals over $\partial D \setminus \overline{\Gamma}$, and using the boundary conditions (2.2)–(2.3), we obtain that the jump $[v]$ satisfies

$$\left(\frac{i}{\lambda}I + T_\Gamma\right)[v] = \frac{i}{\lambda}h + \left(K'_\Gamma - \frac{I}{2}\right)f, \tag{2.10}$$

where the operators $K^{'}_{\Gamma} : \tilde{H}^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ and $T_\Gamma : \tilde{H}^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ are defined by

$$(K'_\Gamma\psi)(x) := \int_\Gamma \psi(y)\frac{\partial}{\partial \nu_x}\mathbb{G}(x,y)\,ds_y \quad \text{for } x \in \Gamma,$$

$$(T_\Gamma\psi)(x) := \frac{\partial}{\partial \nu_x}\int_\Gamma \psi(y)\frac{\partial}{\partial \nu_y}\mathbb{G}(x,y)\,ds_y \quad \text{for } x \in \Gamma,$$

respectively. If (2.10) can be solved for $[v]$, then from the boundary conditions we know $\partial v^+/\partial\nu$ and $\partial v^-/\partial\nu$. Furthermore, it is easy to see that

$$\frac{1}{2}(v^+ + v^-) = -S_\Gamma\left[\frac{\partial u}{\partial \nu}\right] + K_\Gamma[u] \quad \text{on } \Gamma, \tag{2.11}$$

where now $S_\Gamma : \tilde{H}^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ is defined by

$$(S_\Gamma\psi)(x) := \int_\Gamma \psi(y)\,\mathbb{G}(x,y)\,ds_y \quad \text{for } x \in \Gamma.$$

Hence the knowledge of $[v] = v^+ - v^-$ and (2.11) determines v^+ and v^- on Γ and therefore the solution v from the Green representation formula (2.9). To solve (2.10) we observe that $I : \tilde{H}^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ is a compact operator due to Rellich's embedding theorem and that T can be written as a sum of a coercive operator and a compact operator (see Theorems 7.8 and 7.10 in [25]). Hence, since $\lambda(x) > \lambda_0 > 0$, we conclude that $(\frac{i}{\lambda}I + T_\Gamma) : \tilde{H}^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ is a Fredholm operator of index zero. Therefore, it suffices to prove only the injectivity of $\frac{i}{\lambda}I + T_\Gamma$. To this end let $\xi \in \tilde{H}^{1/2}(\Gamma)$ satisfy

$$\left(\frac{i}{\lambda}I + T_\Gamma\right)\xi = 0,$$

and define the following potential:

$$w(x) = \int_\Gamma \xi(y)\frac{\partial}{\partial \nu_y}\mathbb{G}(x,y)ds_y \quad \text{for } x \in \mathbb{R}^2 \setminus \overline{\Gamma}. \tag{2.12}$$

Approaching Γ and using the jump relations for the double layer potential, we obtain

$$\frac{\partial w^+}{\partial \nu} = \frac{\partial}{\partial \nu_x}\int_\Gamma \xi(y)\frac{\partial}{\partial \nu_y}\mathbb{G}(x,y)ds_y = T_\Gamma(\xi),$$

$$[w] = \xi \qquad \text{and} \qquad \left[\frac{\partial w}{\partial \nu}\right] = 0.$$

Hence we have

$$[w] - i\lambda\frac{\partial w^+}{\partial \nu} = \xi - i\lambda T_\Gamma \xi = -i\lambda\left(\frac{i}{\lambda}I + T_\Gamma\right)\xi = 0.$$

Therefore w defined by (2.12) satisfies (2.1)–(2.4) with $f = h = 0$, and from Theorem 2.1, $w = 0$ in $\mathbb{R}^2 \setminus \overline{\Gamma}$, which finally implies $[w] = \xi = 0$. This ends the proof. ◻

3. Reconstruction of the crack from far field data. In this section we assume that the thin dielectric film is embedded in a homogeneous background and the measurements are made from far away. In addition, we assume that the incident field is a time harmonic plane wave given by $u^i := e^{ikx\cdot d}$ for $x \in \mathbb{R}^2$, where the unit vector $d \in S := \{x \in \mathbb{R}^2 : |x| = 1\}$ is the incident direction. In this setting the scattered field u^s satisfies (2.1)–(2.4) with $n(x) = 1$, $f := -\big[\frac{\partial e^{ikx\cdot d}}{\partial \nu}\big] = 0$, and $h := -\big[e^{ikx\cdot d}\big] + i\lambda\frac{\partial e^{ikx\cdot d}}{\partial \nu} = i\lambda\frac{\partial e^{ikx\cdot d}}{\partial \nu}$. Note that $\mathbb{G}(x,y)$ is now the fundamental solution of the Helmholtz equation $\Phi(x,y) := \frac{i}{4}H_0^{(1)}(k|x-y|)$ with $H_0^{(1)}$ being the Hankel function of the first kind of order zero. It is shown in [10] that the scattered field, which now depends also on d, has the asymptotic behavior

$$u^s(x) = \frac{e^{ikr}}{\sqrt{r}}u_\infty(\hat{x}, d) + O(r^{-3/2}), \tag{3.1}$$

where u_∞ is the *far field pattern* of the scattered wave u^s, $\hat{x} = x/|x|$, and $r = |x|$.

The *inverse scattering problem* that we will consider in this section of our paper is to determine Γ and λ from a knowledge of $u_\infty(\hat{x}, d)$ for $\hat{x}$ and d on the unit circle S. Using [5], one can easily generalize the following analysis for the case of limited aperture data, i.e., for $\hat{x}, d \in S_0 \subset S$. For the unique determination of Γ and λ from the above data, see [28] (see also [7], [10]). We will use the *linear sampling method* to solve this inverse problem [6]. To this end, we define the *far field operator* $F : L^2(S) \to L^2(S)$ by

$$(Fg)(\hat{x}) := \int_S u_\infty(\hat{x}, d)g(d)\,ds(d) \tag{3.2}$$

and consider the *far field equation*

$$Fg = \Phi^e_\infty, \tag{3.3}$$

where Φ^e_∞ is the far field pattern of a suitable solution (to be defined later) to the scattering problem. The aim is to characterize the crack Γ by the behavior of an approximate solution g of the far field equation (3.3). We recall that a Herglotz wave function is a solution of the Helmholtz equation in $\mathbb{R}^2$ of the form

$$v_g(x) := \int_S g(d) e^{ikx\cdot d} ds(d), \tag{3.4}$$

where $g \in L^2(S)$ is the *kernel* of v_g. By superposition we have the following relation:

$$(Fg) = \mathcal{B}(i\lambda\mathcal{H}g),$$

where $\mathcal{H} : L^2(S) \to H^{-\frac{1}{2}}(\Gamma)$ is defined by

$$\mathcal{H}g := \frac{\partial v_g}{\partial \nu} \tag{3.5}$$

and $\mathcal{B} : H^{-\frac{1}{2}}(\Gamma) \to L^2(S)$ takes $h \in H^{-\frac{1}{2}}(\Gamma)$ to the far field pattern u_∞ of the solution to (2.1)–(2.4) with $n(x) = 1$, $f := 0$, and h. For $\beta \in \tilde{H}^{\frac{1}{2}}(\Gamma)$ we construct the double layer potential

$$\mathcal{D}(\beta)(x) := \int_\Gamma \beta(y) \frac{\partial}{\partial \nu_y} \Phi(x, y) ds(y),$$

which has as far field pattern $\gamma\mathcal{F}\beta$, where

$$\mathcal{F}\beta := \int_\Gamma \beta(y) \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial \nu_y} ds(y)$$

and $\gamma = \frac{e^{i\pi/4}}{\sqrt{8\pi k}}$. The calculation

$$\int_S g(\hat{y}) \int_\Gamma \beta(x) \frac{\partial}{\partial \nu} e^{-ikx\cdot\hat{y}} ds(x) d(\hat{y}) = \int_\Gamma \beta(x) \int_S g(\hat{y}) \frac{\partial}{\partial \nu} e^{-ikx\cdot\hat{y}} ds(\hat{y}) ds(x)$$

shows that $\mathcal{H}g(-\hat{y})$ is the transpose of $\mathcal{F}\beta$ in the duality pairing between $\tilde{H}^{\frac{1}{2}}(\Gamma)$, $H^{-\frac{1}{2}}(\Gamma)$ and $L^2(S)$, $L^2(S)$, respectively, where $\mathcal{H} : L^2(S) \to H^{-\frac{1}{2}}(\Gamma)$ and $\mathcal{F} : \tilde{H}^{\frac{1}{2}}(\Gamma) \to L^2(S)$.

Lemma 3.1. *The compact operators* $\mathcal{F} : \tilde{H}^{\frac{1}{2}}(\Gamma) \to L^2(S)$ *and* $\mathcal{H} : L^2(S) \to H^{-\frac{1}{2}}(\Gamma)$ *are injective and have dense range, provided that there does not exist a nontrivial Herglotz wave function such that its normal derivative vanishes on* Γ.

Proof. From the above and Lemma 2.10 in [25] it suffices to show that both $\mathcal{F}$ and $\mathcal{H}$ are injective operators. To this end, if $\mathcal{F}(\beta) = 0$, then $\mathcal{D}\beta = 0$ in $\mathbb{R}^2 \setminus \overline{\Gamma}$, which implies $\beta := -[\mathcal{D}\beta] = 0$ from the jump relation. Next, the assumption of the theorem guarantees that $\mathcal{H}$ is also injective. Note that injectivity of $\mathcal{F}$ and consequently the denseness of the range of $\mathcal{H}$ do not require the assumption stated in the lemma. ∎

We remark that, from the above, the far field operator fails to be injective and have dense range if Γ is such that there exists a nontrivial Herglotz wave function with vanishing normal derivative on Γ. An instance of this situation is if Γ is part of a circle of radius r such that kr is a zero of J_1 Bessel function. It is interesting to notice that in the case of the exact model (namely, the obstacle D is a region with nonempty interior), the far field operator fails to be injective and have dense range if the wave number k is a transmission eigenvalue for D with eigenfunction being a Herglotz function (for details on transmission eigenvalues, see [7]). (Note that the injectivity and the denseness of the range of the far field operator are typically needed in most of the inversion schemes in order to use Tikhonov regularization technique.)

From the above analysis and the jump relations applied to the double layer potential $\mathcal{D}$ we see that

$$\mathcal{F}\beta = \gamma^{-1}\mathcal{B}\left(I - i\lambda T_\Gamma\right)\beta,$$

which implies the following factorization of the far field operator:

$$Fg = \gamma\mathcal{F}\left(I - i\lambda T_\Gamma\right)^{-1}\left(i\lambda\mathcal{H}g\right), \qquad g \in L^2(S). \tag{3.6}$$

Lemma 3.2. *For any simple piecewise smooth arc L and $\beta_L \in \tilde{H}^{\frac{1}{2}}(L)$ we define $\Phi^L_\infty \in L^2(S)$ by*

$$\Phi^L_\infty(\hat{x}) := \int_L \beta_L(y)\frac{\partial}{\partial\nu_y}e^{-ik\hat{x}\cdot y}ds_y. \tag{3.7}$$

Then $\Phi^L_\infty(\hat{x}) \in \operatorname{Range}(\mathcal{F})$ if and only if $L \subset \Gamma$.

Proof. First assume that $L \subset \Gamma$. Then since $\tilde{H}^{\frac{1}{2}}(L) \subset \tilde{H}^{\frac{1}{2}}(\Gamma)$ it follows directly from the definition of $\mathcal{F}$ that $\Phi^L_\infty(\hat{x}) \in \operatorname{Range}(\mathcal{F})$.

Now let $L \not\subset \Gamma$ and assume, on the contrary, that $\Phi^L_\infty(\hat{x}) \in \operatorname{Range}(\mathcal{F})$; i.e., there exists $\beta \in \tilde{H}^{\frac{1}{2}}(\Gamma)$ such that

$$\Phi^L_\infty(\hat{x}) = \int_\Gamma \beta(y)\frac{\partial}{\partial\nu_y}e^{-ik\hat{x}\cdot y}ds_y.$$

Hence by Rellich's lemma and the unique continuation principle we have that the potentials

$$\Phi^L(x) = \int_L \beta_L(y)\frac{\partial}{\partial\nu_y}\Phi(x,y)ds_y \qquad \text{and} \qquad \mathcal{D}(x) = \int_\Gamma \beta(y)\frac{\partial}{\partial\nu_y}\Phi(x,y)ds_y$$

coincide in $\mathbb{R}^2 \setminus (\overline{\Gamma} \cup \overline{L})$. Now let $x_0 \in L$, $x_0 \notin \Gamma$, and let $B_\epsilon(x_0)$ be a small ball with center at x_0 such that $B_\epsilon(x_0) \cap \Gamma = \emptyset$. Hence $\mathcal{D}$ is analytic in $B_\epsilon(x_0)$, while Φ^L has a singularity at x_0, which is a contradiction. This proves the lemma. □

Now using Lemma 3.2, the regularization theory for $\mathcal{F}\beta = \Phi^L_\infty$, and the fact that $\left(i\lambda T_\Gamma - I\right)\beta$ can be approximated by $i\lambda\mathcal{H}g$ in $H^{-\frac{1}{2}}(\Gamma)$, we have the following result for the solution of the *far field equation*:

$$(Fg)(\hat{x}) = \gamma\Phi^L_\infty(\hat{x}), \qquad \hat{x} \in S, \tag{3.8}$$

which is the basis of the linear sampling method for reconstructing Γ (cf. Theorem 8.45 of [6]).

Theorem 3.3. *Assume that Γ is a simple piecewise smooth arc and that there does not exist any nontrivial Herglotz wave function having vanishing normal derivative on Γ. Then if F is the far field operator corresponding to* (2.1)–(2.4) *with* $n(x) = 1$,

$f := 0$, and $h := i\lambda \frac{\partial e^{ikx\cdot d}}{\partial \nu}$, the following are true:

1. *If $L \subset \Gamma$, then for every $\epsilon > 0$ there exists a solution $g_\epsilon^L \in L^2(S)$ of the inequality*

$$\|Fg_\epsilon^L - \gamma \Phi_\infty^L\|_{L^2(S)} < \epsilon$$

such that $\mathcal{H}_{g_\epsilon^L}$ converges to a well-defined function in $H^{-\frac{1}{2}}(\Gamma)$.

2. *If $L \not\subset \Gamma$, then for every $\epsilon > 0$ all $g_\epsilon^L \in L^2(S)$ satisfying*

$$\|Fg_\epsilon^L - \gamma \Phi_\infty^L\|_{L^2(S)} < \epsilon$$

are such that

$$\lim_{\epsilon \to 0} \|g_\epsilon^L\|_{L^2(S)} = \infty \quad \textit{and} \quad \lim_{\epsilon \to 0} \|\mathcal{H}_{g_\epsilon^L}\|_{H^{-\frac{1}{2}}(\Gamma)} = \infty,$$

where $\mathcal{H}_{g_\epsilon^L}$ is defined by (3.5) with $v_{g_\epsilon^L}$ being the Herglotz wave function with kernel g_ϵ^L.

Remark 3.1. From the above analysis we notice that, for $L \subset \Gamma$, $i\lambda\mathcal{H}_{g_\epsilon^L}$ approximates $(I - i\lambda T_\Gamma)\beta_L$, where g_ϵ^L is the approximate solution of (3.8). In particular, in the $H^{-\frac{1}{2}}(\Gamma)$-norm we have

$$\frac{\partial v_{g_\epsilon^L}}{\partial \nu} \approx \frac{i\beta_L}{\lambda} - T_\Gamma \beta_L, \qquad L \subset \Gamma \quad \text{and} \quad \beta_L \in C_0^\infty(L), \tag{3.9}$$

which can be used to recover λ, provided that (a reconstruction of) Γ is now known (e.g., by using the linear sampling method based on Theorem 3.3).

4. Reconstruction of the crack from near field data. We now assume that the dielectric thin film is embedded in a known inhomogeneous medium with index of refraction $n(x)$ that satisfies the assumptions stated in section 2. The incident field is a point source given by $\Phi(x, x_0, k_s) := \frac{i}{4}H_0^{(1)}(k_s|x - x_0|)$ located at a point x_0 outside a bounded region Ω surrounding the crack and $k_s^2 = k^2 n(x_0)$ (see Figure 4.1). In this case the scattered field u^s satisfies (2.1)–(2.4) with $f := -\left[\frac{\partial\Phi(\cdot,x_0,k_s)}{\partial\nu}\right] = 0$ and $h := -\left[\Phi(\cdot, x_0, k_s)\right] + i\lambda\frac{\partial\Phi(\cdot,x_0,k_s)}{\partial\nu} = i\lambda\frac{\partial\Phi(\cdot,x_0,k_s)}{\partial\nu}$. Note that u^s is the sum of the scattered field due to the crack and the scattered field due to the medium. Let $u(\cdot, x_0) = u^s(\cdot, x_0) + \Phi(\cdot, x_0, k_s)$ denote the total field, let Λ be a closed curve containing Ω, and suppose that in a neighborhood of Λ the index of refraction $n(x)$ is constant and $k^2 n = k_s^2$. Note that Λ can be part of a closed analytic curve, and by an analycity argument the following analysis holds true as well. For technical reasons we write $u(\cdot, x_0) = u_c^s(\cdot, x_0) + \mathbb{G}(\cdot, x_0)$, where $u_c^s(\cdot, x_0)$ is the scattered field due to the crack and $\mathbb{G}(\cdot, x_0)$ is the total field due to the medium or the background Green's function satisfying (2.8).

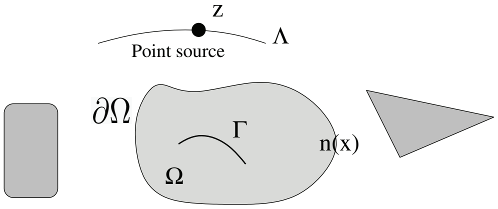


FIG. 4.1. *Geometry for the reciprocity gap functional method.*

The *inverse scattering problem* we are interested in now is to determine the crack Γ from a knowledge of $u(\cdot, x_0)$ and $\frac{\partial u(\cdot, x_0)}{\partial \nu}$ on the boundary $\partial\Omega$ for all point sources located at any $x_0 \in \Lambda$. These measurements in the case of Maxwell's equations correspond to measuring the tangential components of the electric and magnetic fields. To solve this inverse scattering problem we adapt the *reciprocity gap functional* method first introduced [12] for obstacles with nonempty interior. To this end, let

$$\mathbb{H}(\Omega) := \left\{ w \in H^1(\Omega) : \ \Delta w + k^2 n(x) w = 0 \quad \text{in } \Omega \right\}.$$

In a similar way as in Lemma 3.1 (see also the proof of Lemma 4.4 in [8]) it can be shown that the set

$$\left\{ (\mathcal{S}\varphi)(y) := \int_\Sigma \varphi(x)\Phi(x,y)\, ds_x \quad \text{for } \varphi \in L^2(\Sigma) \right\} \tag{4.1}$$

is dense in $\mathbb{H}(\Omega)$, where Σ is a open curve outside Ω and $\Phi(x,y)$ is the radiating fundamental solution of $\Delta u + k^2 \tilde{n}(x) u = 0$, where $\tilde{n}(x) = n(x)$ for $x \in \Omega$ and $\tilde{n}(x) = 1$ for $x \in \mathbb{R}^2 \setminus \Omega$ (or any other convenient extension). In particular, if $n(x) = n_0$ is constant in Ω, then $\Phi(x,y)$ is simply $\frac{i}{4} H_0^{(1)}(k\sqrt{n_0}|x - x_0|)$.

We define the *reciprocity gap operator* $\mathcal{R} : \mathbb{H}(\Omega) \to L^2(\Lambda)$ by

$$\mathcal{R}(w)(x_0) = \int_{\partial\Omega} \left(u(\cdot, x_0) \frac{\partial w}{\partial \nu} - w \frac{\partial u(\cdot, x_0)}{\partial \nu} \right) ds, \qquad x_0 \in \Lambda, \ w \in \mathbb{H}(\Omega). \tag{4.2}$$

THEOREM 4.1. *The compact operator $\mathcal{R} : \mathbb{H}(\Omega) \to L^2(\Lambda)$ is injective and has dense range, provided that there does not exist any $w \in \mathbb{H}(\Omega)$ such that $\partial w / \partial \nu = 0$ on Γ.*

Proof. If $\mathcal{R}(w) = 0$, applying Green's second identity and using the zero boundary condition for $u(\cdot, x_0)$, it is easy to see that

$$0 = \mathcal{R}(w)(x_0) = -i \int_\Gamma \lambda \frac{\partial u(\cdot, x_0)}{\partial \nu} \frac{\partial w}{\partial \nu} ds.$$

Noting that from the boundary condition $\lambda \partial u(\cdot, x_0)/\partial \nu \in \tilde{H}^{\frac{1}{2}}(\Gamma)$ it suffices to show that $\lambda \partial u(\cdot, x_0)/\partial \nu$ for $x_0 \in \Lambda$ are dense in $\tilde{H}^{\frac{1}{2}}(\Gamma)$. Indeed, this fact implies that $\partial w / \partial \nu = 0$, which contradicts the assumption of the theorem. To prove the denseness property let $\psi \in H^{-\frac{1}{2}}(\Gamma)$ such that

$$0 = \int_\Gamma \lambda(x) \psi(x) \frac{\partial u(x, x_0)}{\partial \nu}\, ds_x \qquad \text{for all } x_0 \in \Lambda,$$

where the integral is understood in the sense of duality pairing, and let $w \in H^1(\mathbb{R}^2 \setminus \overline{\Gamma})$ be the solution of (2.1)–(2.4) with $f := 0$ and $h := \lambda\psi$. Applying Green's second identity to w and u_c^s (note that both satisfy $\Delta u + k^2 n u = 0$ outside Γ) and the boundary condition for $u(\cdot, x_0)$, we obtain that

$$\begin{aligned}
0 &= \int_\Gamma \lambda(x)\psi(x)\frac{\partial u(x,x_0)}{\partial \nu}\, ds_x = \int_\Gamma \left([w] - i\lambda\frac{\partial w}{\partial \nu}\right)\frac{\partial\left(u_c^s + \mathbb{G}(\cdot,x_0)\right)}{\partial \nu}\, ds \\
&= \int_\Gamma \left([u_c^s] - i\lambda\frac{\partial u_c^s}{\partial \nu}\right)\frac{\partial w}{\partial \nu}\, ds + \int_\Gamma \left([w] - i\lambda\frac{\partial w}{\partial \nu}\right)\frac{\partial \mathbb{G}(\cdot,x_0)}{\partial \nu}\, ds \\
&= -\int_\Gamma \left([\mathbb{G}(\cdot,x_0)] + i\lambda\frac{\partial \mathbb{G}(\cdot,x_0)}{\partial \nu}\right)\frac{\partial w}{\partial \nu}\, ds + \int_\Gamma \left([w] - i\lambda\frac{\partial w}{\partial \nu}\right)\frac{\partial \mathbb{G}(\cdot,x_0)}{\partial \nu}\, ds \\
&= \int_\Gamma [w]\frac{\partial \mathbb{G}(\cdot,x_0)}{\partial \nu}\, ds \qquad \text{for all } x_0 \in \Lambda.
\end{aligned}$$

Hence $P(x_0) = \int_\Gamma [w(x)]\frac{\partial \mathbb{G}(x,x_0)}{\partial \nu}\, ds_x$ is a radiating solution as a function of x_0 which vanishes on Λ, which implies that $P(x_0) = 0$ outside the domain bounded by Λ and consequently in $\mathbb{R}^2 \setminus \overline{\Gamma}$ by unique continuation. Hence, using the jump relations, we have that $[w] = 0$ on Γ, which together with $[\partial w/\partial \nu] = 0$ on Γ implies $w = 0$ and consequently $\psi = 0$.

Next we show that $\mathcal{R}$ has dense range. Let $\alpha \in L^2(\Lambda)$ be such that $(\mathcal{R}w, \overline{\alpha})_{L^2(\Lambda)} = 0$ for all $w \in \mathbb{H}(\Omega)$. The bilinearity of $\mathcal{R}$ implies that

$$(4.3) \qquad (\mathcal{R}w, \overline{\alpha})_{L^2(\Lambda)} = \int_{\partial\Omega}\left(Q\frac{\partial w}{\partial \nu} - w\frac{\partial Q}{\partial \nu}\right) ds = -i\int_\Gamma \lambda\frac{\partial Q}{\partial \nu}\frac{\partial w}{\partial \nu} ds = 0$$

for all $w \in \mathbb{H}(\Omega)$, where

$$Q(x) = \int_\Lambda \alpha(x_0)u(x,x_0)\, ds(x_0) = \int_\Lambda \alpha(x_0)u_c^s(x,x_0)\, ds(x_0) + \int_\Lambda \alpha(x_0)\mathbb{G}(x,x_0)\, ds(x_0).$$

Hence (4.3) implies that $\partial Q/\partial \nu = 0$ on Γ since obviously the set $\{\partial w/\partial \nu :\ w \in \mathbb{H}(\Omega)\}$ is dense in $H^{-\frac{1}{2}}(\Gamma)$ and Q is smooth near Γ. Since $Q \in \mathbb{H}(\Omega)$, from the assumption we can conclude that $Q = 0$ in Ω and therefore by unique continuation in the domain bounded by Λ. Since Q is continuous across Λ we conclude that Q is a radiating solution and is zero on Λ which implies that $Q = 0$. Finally by the jump relation of the normal derivative of single layer potential we finally have that $\alpha = 0$, which ends the proof. □

Now we have all the ingredients to describe a sampling algorithm to determine Γ without knowing λ. Let L be an open arc in Ω and consider

$$\Phi^L(x) = \int_L \beta_L(y)\frac{\partial}{\partial \nu_y}\Phi(x,y)ds_y, \qquad \beta_L \in \tilde{H}^{\frac{1}{2}}(L),$$

where $\Phi(x,y)$ satisfies

$$\Delta\Phi(x,y) + k^2\tilde{n}(x)\Phi(x,y) = \delta(x,y)$$

and again $\tilde{n}(x) = n(x)$ for $x \in \Omega$ and $\tilde{n}(x) = 1$ for $x \in \mathbb{R}^2 \setminus \Omega$ (or any other convenient extension). Note that we need to know only the index of refraction of the background medium inside Ω, and this is the strength of this method compared to the linear sampling method. Again if $n(x) = n_0$ is constant in Ω, one can choose $\Phi(x,y) := \frac{i}{4}H_0^{(1)}(k\sqrt{n_0}|x-x_0|)$. The proposed algorithm consists of seeking for each open arc $L \subset \Omega$ a solution $\varphi \in L^2(\Sigma)$ to the first kind ill-posed linear equation

$$\mathcal{R}(\mathcal{S}\varphi)(x_0) = \mathcal{R}(\Phi^L)(x_0), \qquad x_0 \in \Lambda. \tag{4.4}$$

Note that in order to guarantee that $\mathcal{RS}$ is injective and has dense range from the proof of Theorem 4.1 it suffices to assume that there does not exist any potential in (4.1) with vanishing normal derivative on Γ. This condition is similar to the condition we imposed on the Herglotz functions in section 3, and it excludes some special types of cracks.

Note also that in (4.4), $\mathcal{S}\varphi$ can be replaced with any one parameter dense family of functions in $\mathbb{H}(\Omega)$. We refer the reader to [26] for a variational approach to constructing such a family. In particular, if $n(x)$ is constant in Ω, one could use the corresponding Herglotz wave functions using the result of [11].

Now, if $L \subset \Gamma$, from the proof of the first part of Theorem 4.1 we see that (4.4) has a unique solution if and only if $\frac{\partial \mathcal{S}\varphi}{\partial \nu} = \frac{\partial \Phi^L}{\partial \nu}$ on Γ. This can only be satisfied approximately for some $\varphi \in L^2(\Sigma)$ since the set $\left\{\partial\mathcal{S}\varphi/\partial\nu : \varphi \in L^2(\Sigma)\right\}$ is dense in $H^{-\frac{1}{2}}(\Gamma)$.

Next, if $L \not\subset \Gamma$, we can find φ_ϵ such that

$$\|\mathcal{R}(\mathcal{S}\varphi_\epsilon) - \mathcal{R}(\Phi^L)\|_{L^2(\Lambda)} < \epsilon$$

and $\|\partial\mathcal{S}\varphi_\epsilon/\partial\nu\|_{H^{-\frac{1}{2}}(\Gamma)} < C$. Since $u(\cdot,x_0) = u_c^s(\cdot,x_0) + \mathbb{G}(\cdot,x_0)$, using Green's formula we obtain

$$\begin{aligned}
\mathcal{R}(\Phi^L)(x_0) &= \int_{\partial\Omega}\left(u_c^s(x,x_0)\frac{\partial\Phi^L(x)}{\partial\nu} - \Phi_L(x)\frac{\partial u_c^s(x,x_0)}{\partial\nu}\right)ds_x \\
&\quad + \int_{\partial\Omega}\left(\mathbb{G}(x,x_0)\frac{\partial\Phi^L(x)}{\partial\nu} - \Phi_L(x)\frac{\partial\mathbb{G}(x,x_0)}{\partial\nu}\right)ds_x \\
&= w(x_0) + \int_L \beta_L(y)\frac{\partial}{\partial\nu_y}\int_{\partial\Omega}\left(\mathbb{G}(x,x_0)\frac{\partial\Phi(x,y)}{\partial\nu} - \Phi(x,y)\frac{\partial\mathbb{G}(x,x_0)}{\partial\nu}\right)ds_x\,ds_y \\
&= w(x_0) + \int_L \beta_L(y)\frac{\partial}{\partial\nu_y}\mathbb{G}(x_0,y)ds_y,
\end{aligned}$$

where $w(x_0)$ is a solution to $\Delta_{x_0}w + k^2 n(x_0)w = 0$. On the other hand,

$$\mathcal{R}(\mathcal{S}\varphi_\epsilon)(x_0) = -i\int_\Gamma \lambda\frac{\partial u(x,x_0)}{\partial\nu}\frac{\partial\mathcal{S}\varphi_\epsilon}{\partial\nu}ds.$$

Since $\|\partial\mathcal{S}\varphi_\epsilon/\partial\nu\|_{H^{-\frac{1}{2}}(\Gamma)} < C$, we can assume that there exists a sequence such that $\lim_{\epsilon\to 0}\partial\mathcal{S}\varphi_\epsilon/\partial\nu = \theta \in H^{-\frac{1}{2}}(\Gamma)$ weakly, whence

$$\lim_{\epsilon\to 0}\mathcal{R}(\mathcal{S}\varphi_\epsilon)(x_0) = -i\int_\Gamma \lambda\frac{\partial u(x,x_0)}{\partial\nu}\,\theta(x)ds_x.$$

Hence

$$-i\int_\Gamma \lambda \frac{\partial u_c^s(x,x_0)}{\partial \nu}\,\theta(x)ds_x - i\int_\Gamma \lambda \frac{\partial \mathbb{G}(x,x_0)}{\partial \nu}\,\theta(x)ds_x$$

$$= w(x_0) + \int_L \beta_L(x)\frac{\partial}{\partial \nu_x}\mathbb{G}(x_0,x)ds_x. \tag{4.5}$$

Since the first term on both sides can be extended as solution to $\Delta_{x_0}w+k^2n(x_0)w=0$ outside the domain bounded by Λ, we deduce by uniqueness and the unique continuation principle that (4.5) holds in $\mathbb{R}^2\setminus\overline{\Gamma}\cup\overline{L}$. Now we arrive at a contradiction since for $x_0\in L$, $x_0\notin\Gamma$, and $B_\epsilon(x_0)$ a small ball with center at x_0 such that $B_\epsilon(x_0)\cap\Gamma=\emptyset$ the left-hand side is analytic whereas the right-hand side has a singularity at x_0.

The above analysis has proven the following main theorem of this section, which is the basis of the linear sampling method based on the reciprocity gap functional for determining Γ.

Theorem 4.2. *Assume that* Γ *is simple piecewise smooth arc and that there does not exist any potential in* (4.1) *having zero normal derivative on* Γ. *Then if* $u(\cdot,x_0)$ *is the total field corresponding to* (2.1)–(2.4) *with* $f:=0$ *and* $h:=i\lambda\frac{\partial\Phi(\cdot,x_0,k_s)}{\partial\nu}$, *the following are true:*

1. *If* $L\subset\Gamma$, *then for every* $\epsilon>0$ *there exists a* $\varphi_\epsilon^L\in L^2(\Sigma)$ *satisfying*

$$\|\mathcal{R}(\mathcal{S}\varphi_\epsilon^L)-\mathcal{R}(\Phi^L)\|_{L^2(\Lambda)}<\epsilon$$

such that $\frac{\partial\mathcal{S}\varphi_\epsilon^L}{\partial\nu}$ *converges to* $\frac{\partial\Phi^L}{\partial\nu}$ *in* $H^{-\frac{1}{2}}(\Gamma)$.

2. *If* $L\not\subset\Gamma$, *then for every* $\epsilon>0$ *any* $\varphi_\epsilon^L\in L^2(\Sigma)$ *satisfying*

$$\|\mathcal{R}(\mathcal{S}\varphi_\epsilon^L)-\mathcal{R}(\Phi^L)\|_{L^2(\Lambda)}<\epsilon$$

is such that

$$\lim_{\epsilon\to0}\|\varphi_\epsilon^L\|_{L^2(\Sigma)}=\infty \quad \textit{and} \quad \lim_{\epsilon\to0}\left\|\frac{\partial\mathcal{S}\varphi_\epsilon^L}{\partial\nu}\right\|_{H^{-\frac{1}{2}}(\Gamma)}=\infty,$$

where $\mathcal{S}\varphi_\epsilon^L$ *is defined by* (4.1).

5. Numerical examples.

5.1. The linear sampling method.

In this section we will give some results of numerical experiments for identifying cracks based on the theory developed in section 3. The far field data we use are synthetic, but corrupted by random noise added pointwise to the measurements. The forward problem is numerically solved using a quadrature method for the first kind hypersingular integral equation (2.10) as developed by Kress and co-authors in [9], [19], and [20]. This method seems well suited to our problem since we need to invert a hypersingular integral operator. This claim is validated by our numerical results which show a convergence rate similar to that in [9]. However, the exact order of singularity of the solution of the forward problem at the tips of the crack still remains to be studied. The computed far field data is obtained as a trigonometric series $u_\infty=\sum_{n=-N}^{N}u_{\infty,n}\exp(in\theta)$. We then add random noise to the Fourier coefficients of u_∞ to obtain the approximate far field pattern $u_{\infty,a}=\sum_{n=-N}^{N}u_{\infty,a,n}\exp(in\theta)$, where $u_{\infty,a,n}=u_{\infty,n}(1+\epsilon\chi_n)$ with χ_n a random variable in $[-1,\ 1]$ ($\epsilon=0.05$ in our examples). We remark that this random noise is rather "special," since the far field data polluted in this way remains the far field of some radiating solution to the Helmholtz equation, as the decay rate of the Fourier coefficients is not modified. This may not be the case for the noise in measured data. Also, it would be interesting to test our inversion method using simulated data computed from the exact model of the forward problem. Unfortunately, we do not have available a forward code to do so. Notice that we do not commit any inverse crime since the method for solving the direct problem and the method for solving the inverse problem are completely different.

The inversion scheme is based on solving the following ill-posed first kind equation:

$$\int_S u_\infty(\hat{x}, d) g(d)\, ds(d) = \gamma \int_L \beta_L(y) \frac{\partial}{\partial \nu_y} e^{-ik\hat{x}\cdot y} ds_y \tag{5.1}$$

for an arbitrary open arc L and $\beta_L \in \tilde{H}^{\frac{1}{2}}(L)$. Then the crack can be reconstructed by using the fact that if $L \subset \Gamma$, we can find a bounded $g \in L^2(S)$ that satisfies (5.1) with discrepancy ϵ, whereas if $L \not\subset \Gamma$, all approximate solutions to (5.1) are unbounded. In this study we search for the crack by taking L to be a small segment centered at a sampling point z with unit normal vector n_z and β_L a sequence that converges to $\delta(z)$. Thus, in the limiting case, (5.1) is replaced by

$$\int_{-\pi}^{\pi} u_\infty(\hat{x}, \theta) g_{z,n_z}(\theta)\, d\theta = -ik\gamma\, n_z \cdot \hat{x}\, e^{-ik\hat{x}\cdot z}, \qquad \hat{x} \in S,\ z \in \mathbb{R}^2,\ n_z \in S. \tag{5.2}$$

Now, if $z \in \Gamma$ and n_z coincides with the normal vector to Γ at z, then we can find a bounded $g_{z,n_z} \in L^2(S)$ that approximately solves (5.2). Otherwise all such $g_{z,n_z} \in L^2(S)$ are unbounded.

In order to solve (5.2) we use Tikhonov regularization and the Morozov discrepancy principle to deal with the severe ill-posedness of this equation. In particular, using the above expression for $u_{\infty,a}$, (5.2) is rewritten as an ill-conditioned matrix equation for the Fourier coefficients of g, which we write in the form

$$A g_{z,\theta_z} = f_{z,\theta_z}, \qquad \text{where } \theta_z = n_z \cdot \hat{x} \in [-\pi, \pi]. \tag{5.3}$$

As already noted, this equation needs to be regularized. To do this, we begin by computing the singular value decomposition of A, i.e., $A = U\Lambda V^*$, where U and V are unitary and Λ is real diagonal with $\Lambda_{l,l} = \sigma_l,\ 1 \le l \le n$, where σ_l are the singular values of A. The solution of (5.3) is then equivalent to solving $\Lambda V^* g_{z,\theta_z} = U^* f_{z,\theta_z}$. Now letting $\rho_{z,\theta_z} = (\rho_1, \rho_2, \ldots, \rho_n)^\top = U^* f_{z,\theta_z}$, the Tikhonov regularization of (5.3) leads to the problem of solving

$$\min_{g_{z,\theta_z} \in \mathbb{R}^n} \|\Lambda V^* g_{z,\theta_z} - \rho_{z,\theta_z}\|_{l^2}^2 + \alpha \|g_{z,\theta_z}\|_{l^2}^2,$$

where $\alpha > 0$ is the regularization parameter [6]. The numerical procedure for locating the crack is the following: we consider a uniform grid of sampling points $\{z_i\}_{i=1,\ldots,N}$ in the probing area, and for each sampling point z_i we choose a finite number of angles $\theta_{z_i}^j$, $j = 1, \ldots, M$, and compute

$$\mathcal{G}(z_i, \theta_{z_i}^j) = \|f_{z_i,\theta_{z_i}^j}\|_{\ell^2} / \|g_{z,\theta_{z_i}^j}\|_{\ell^2}.$$

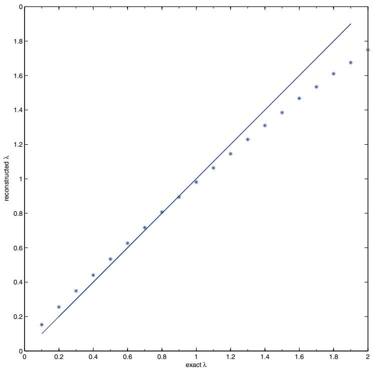


Fig. 5.1. *Here we show the reconstruction of different values of* λ *for the curved crack shown in Figure* 5.3(a). *The reconstructed values are shown by the little stars. The true reconstruction would be on the solid line.*

It is expected that $\mathcal{G}(z_i, \theta^j_{z_i})$ becomes relatively large if z_i is a point of Γ and $\theta^j_{z_i}$ is (near) the angle corresponding to the tangent line to Γ at z_i. In all our reconstruction examples we plot the indicator function

$$\mathbb{G}(z_i) = \sum_{j=1}^{M} \mathcal{G}(z_i, \theta^j_{z_i}) \qquad \text{for all sampling points } z_i \text{ on the grid.}$$

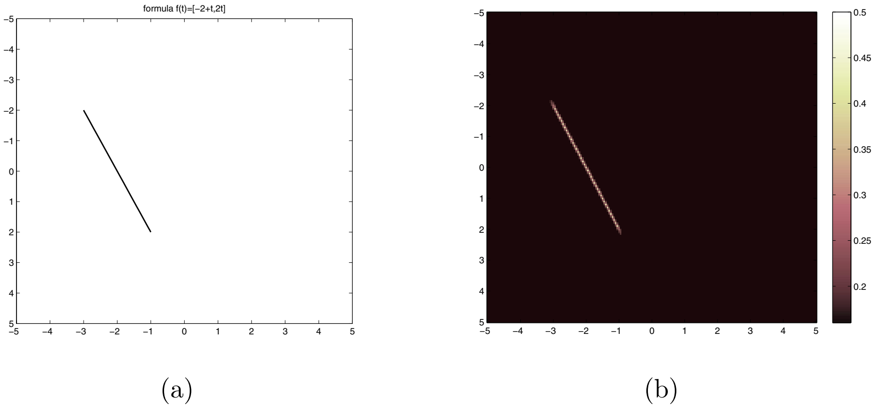
We found out that one obtains similar plots when taking the maximum over all $\theta^j_{z_i}$ for every fixed z_i instead of the summation. However, further investigation is needed to construct an indicator function that better captures the effect of the angles θ_z.

Having reconstructed Γ, it is possible to use (3.9) to reconstruct λ. In this work we have not investigated the best numerical strategy to implement (3.9) in the case when λ is a function. Here we present some preliminary results in the simplest case when λ is a constant. To this end it is natural to consider the imaginary part of (3.9). Hence the reconstruction formula is based on

$$\operatorname{Im} \frac{\partial v_{g^L_\epsilon}}{\partial \nu}(x) = \frac{1}{\lambda}\beta_L(x) - \operatorname{Im}(T_\Gamma \beta_L)(x), \qquad x \in \Gamma,$$

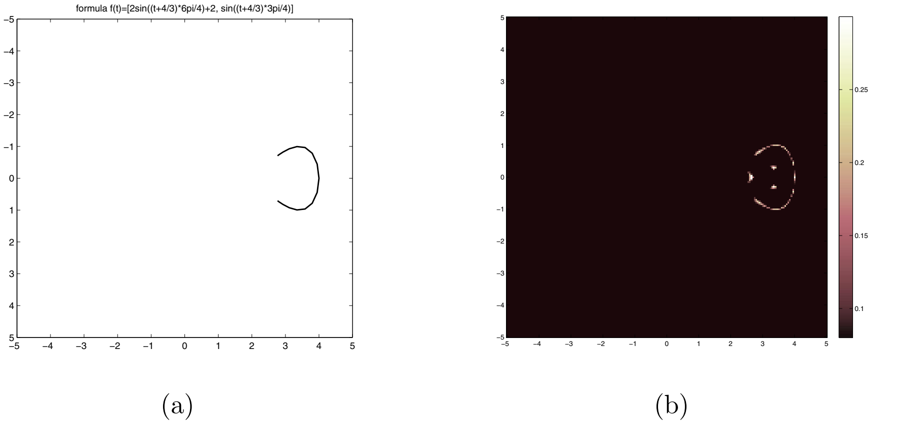
where $\beta_L \in C^\infty_0(L)$. (Note that the imaginary part of the potential T_Γ is an operator with a smooth kernel.) Now if λ is constant, we fix a point in $z \in \Gamma$ and the normal vector n_z to the crack at z and let g_{z,n_z} be the corresponding solution of the discrete far field equation. Then $\operatorname{Im}\frac{\partial v_{g_{z,n_z}}(z)}{\partial n_z}$ is approximately $1/\lambda A_z + B_z$, where A_z and B_z do not depend on λ. Hence, it is possible to avoid unstable computation of A_z and B_z, by computing the Herglotz wave function for two values of λ. Our preliminary examples show that the determination of A_z and B_z is robust. An example of reconstruction of λ based on this approach is shown in Figure 5.1. However, we note that more numerical study is needed to make the reconstruction formula for λ more practical especially for nonconstant λ. The information on λ is useful since it contains knowledge about the thickness of the dielectric object as well as its physical properties.

The numerical examples presented here consist of using the linear sampling method to reconstruct the shape of the dielectric crack, as explained above, for the following



(a) (b)

FIG. 5.2. *Panel* (a) *shows the exact crack and panel* (b) *the reconstruction using the linear sampling method. The wave number is* $k = 3$.



(a) (b)

FIG. 5.3. *Panel* (a) *shows the exact crack and panel* (b) *the reconstruction using the linear sampling method. The wave number is* $k = 3$.

open arcs:

$$\Gamma := \{-2 + s, 2s : -1 \leq s \leq 1\},$$

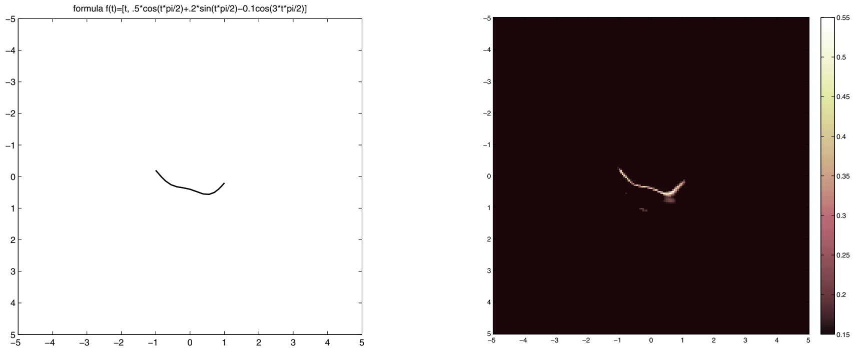
shown in Figure 5.2(a);

$$\Gamma := \left\{2\sin\left(\frac{3\pi}{2}s\right) + 2, \sin\left(\frac{3\pi}{2}s + \pi\right) : -1 \leq s \leq 1\right\},$$

shown in Figure 5.3(a); and

$$\Gamma := \left\{s, 0.5\cos\frac{\pi s}{2} + 0.2\sin\frac{\pi s}{2} - 0.1\cos\frac{3\pi s}{2} : -1 \leq s \leq 1\right\},$$

shown in Figure 5.4(a). The respective reconstructions are shown in Figures 5.2(b), 5.3(b), and 5.4(b). In all reconstructions we keep $k = 3$ and the noise level 5%.



(a) (b)

FIG. 5.4. *Panel* (a) *shows the exact crack and panel* (b) *the reconstruction using the linear sampling method. The wave number is* $k = 3$.

5.2. The reciprocity gap functional method. Here we assume that $n(x) = n_0$ for $x \in \overline{\Omega}$, where n_0 is a complex constant and $n(x) = 1$ outside Ω. Similarly to the case of the linear sampling method, we take L to be a small segment centered at a sampling point z with unit normal vector n_z and β_L a sequence converging to $\delta(z)$. Hence, in the same manner as for the linear sampling method, we can write (4.4) as

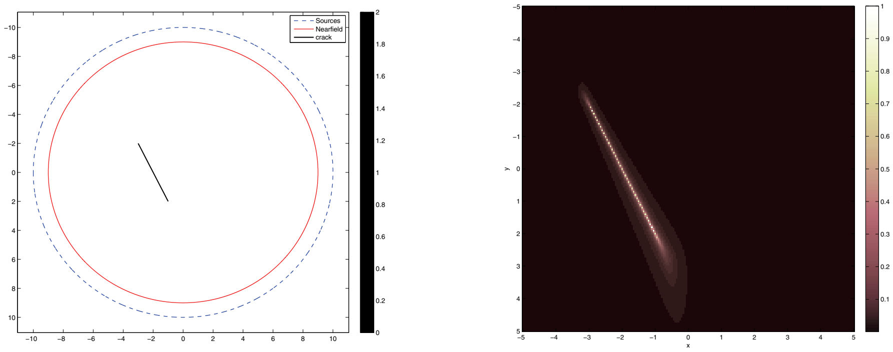
$$A\varphi_{z,n_z}(x_0) = f_{z,n_z}(x_0), \qquad x_0 \in \Lambda, \tag{5.4}$$

where $A : L^2(\Sigma) \to L^2(\Lambda)$ is the integral operator with kernel

$$K(x, x_0) := \mathcal{R}(H_0^{(1)}(k\sqrt{n_0}|x - (\cdot)|))(x_0)$$

and $f_{z,n_z}(x_0) = \mathcal{R}(ikn_z \cdot \nabla H_0^{(1)}(k\sqrt{n_0}|z - (\cdot)|))(x_0)$. Equation (5.4) is an ill-posed linear integral equation and is solved in the same way as explained in section 5.1. Similarly, the approximate solution φ_{z,n_z} is then used to identify the crack.

We end by showing an example of reconstruction for a linear crack embedded in a homogeneous medium surrounded by a circle with $n_0 = 1.5$. The data are measured on the upper half of a bigger circle. We are limited here to small contrast for the host medium, due only to the lack of forward data; it is not a limitation of the reciprocity gap functional method. We computed the data by adding the scattered field due to a crack embedded in a homogeneous media with $n = 1$ in $\mathbb{R}^2$ and the scattered field due to the disk with $n = 1.9$; i.e., we are ignoring multiple scattering effects. The example presented in Figure 5.5 indicates reliable performance of the RGF (reciprocity gap functionals) method based on sampling. Certainly, more numerical experiments are needed to validate it, including the case of absorbing background and limited aperture measurements.



(a) (b)

Fig. 5.5. *Panel* (a) *shows the configuration of the example. The crack is the black solid line embedded inside the red solid circle $\partial\Omega$ corresponding to $n_0 = 1.9$. The sources are paced on the upper half of the blue dashed circle denoted by Σ. The measurements are made on the red circle $\partial\Omega$. Panel* (b) *shows the reconstructed crack using the reciprocity gap functional method. A zoom of the area inside the red circle containing the crack is shown. The wave number is $k = 5$.*

Acknowledgment. This work is part of the Ph.D. Dissertation of N. Zeev under the supervision of F. Cakoni.

REFERENCES

[1] H. Ammari, H. Kang, and F. Santosa, *Scattering of electromagnetic waves by thin dielectric planar structures*, SIAM J. Math. Anal., 38 (2006), pp. 1329–1342.

[2] A. Ben Abda, F. Delbary, and H. Haddar, *On the use of the reciprocity-gap functional in inverse scattering from planar cracks*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1553–1574.

[3] J. R. Bowler, *Thin skin eddy-current inversion for the determination of crack shapes. Special section on electromagnetic and ultrasonic nondestructive evaluation*, Inverse Problems, 18 (2002), pp. 1891–1905.

[4] M. Brühl, M. Hanke, and M. Pidcock, *Crack detection using electrostatic measurements*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 595–605.

[5] F. Cakoni and D. Colton, *Combined far field operators in electromagnetic inverse scattering theory*, Math. Methods Appl. Sci., 26 (2003), pp. 413–429.

[6] F. Cakoni and D. Colton, *The linear sampling method for cracks*, Inverse Problems, 19 (2003), pp. 279–295.

[7] F. Cakoni and D. Colton, *Qualitative Methods in Inverse Scattering Theory*, Springer, Berlin, 2006.

[8] F. Cakoni, F. M. B. Fares, and H. Haddar, *Analysis of two linear sampling methods applied to electromagnetic imagining of buried objects*, Inverse Problems, 22 (2006), pp. 845–867.

[9] R. Chapko, R. Kress, and L. Mönch, *On the numerical solution of a hypersingular integral equation for elastic scattering from a planar crack*, IMA J. Numer. Anal., 20 (2000), pp. 601–619.

[10] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.

[11] D. Colton and R. Kress, *On the denseness of Herglotz wave functions and electromagnetic Herglotz pairs in Sobolev spaces*, Math. Methods Appl. Sci., 24 (2001), pp. 1289–1303.

[12] D. Colton and H. Haddar, *An application of the reciprocity gap functional to inverse scattering theory*, Inverse Problems, 21 (2005), pp. 383–398.

[13] D. Dobson and F. Santosa, *Nondestructive evaluation of plates using eddy current methods*, Internat. J. Engrg. Sci., 36 (1998), pp. 395–409.

[14] S. Fan, J. Winn, A. Devrnyi, J. Chen, R. Meade, and J. Joannopoulos, *Guided and defected modes in periodic waveguides*, J. Opt. Soc. Amer. B Opt. Phys., 12 (1995), pp. 1267–1283.

[15] H. Haddar, *Interface Conditions for Thin Dielectric Layers*, preprint.

[16] H. Haddar, P. Joly, and H. M. Nguyen, *Generalized impedance boundary conditions for scattering by strongly absorbing obstacles: The scalar case*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1273–1300.

[17] O. Ivanyshyn and R. Kress, *Nonlinear integral equations for solving inverse boundary value problems for inclusions and cracks*, J. Integral Equations Appl., 18 (2006), pp. 13–38.

[18] A. Kirsch and S. Ritter, *A linear sampling method for inverse scattering from an open arc*, Inverse Problems, 16 (2000), pp. 89–105.

[19] R. Kress, *Linear Integral Equations*, 2nd ed., Springer, New York, 1999.

[20] R. Kress, *On the numerical solution of a hypersingular integral equation in scattering theory*, J. Comput. Appl. Math., 61 (1995), pp. 345–360.

[21] R. Kress, *Inverse scattering from an open arc*, Math. Methods Appl. Sci., 18 (1995), pp. 267–293.

[22] D. Lesselier and J. Bowler, eds., *Special section on electromagnetic and ultrasonic nondestructive evaluation*, Inverse Problems, 18 (2002), pp. 1733–1963.

[23] J. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York, Heidelberg, Berlin, 1972.

[24] M. McIver, *An inverse problem in electromagnetic crack detection*, IMA J. Appl. Math., 47 (1991), pp. 127–145.

[25] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[26] P. Monk and J. Sun, *Inverse scattering using finite elements and gap reciprocity for inhomogeneous media*, Inverse Problems and Imaging, 1 (2007), pp. 643–660.

[27] S. Moskow, F. Santosa, and J. Zhang, *An approximate method for scattering by thin structures*, SIAM J. Appl. Math., 66 (2005), pp. 187–205.

[28] N. Zeev, *Direct and Inverse Scattering Problems for Thin Obstacles and Interfaces*, Doctoral thesis, Department of Mathematical Sciences, University of Delaware, Newark, DE, 2008.

[29] A. Tamburrino, M. Morozov, G. Rubinacci, and S. Ventres, *Numerical models of volumetric insulating cracks in eddy-current testing with experimental validation*, IEEE Trans. Magnetics, 42 (2006), pp. 1568–1576.

SIAM J. APPL. MATH.
Vol. 69, No. 4, pp. 1043–1064

MODELING, SIMULATION, AND DESIGN FOR A CUSTOMIZABLE ELECTRODEPOSITION PROCESS[*]

PRADEEP THIYANARATNAM[†], RUSSEL CAFLISCH[‡], PAULO S. MOTTA[§], AND JACK W. JUDY[¶]

Abstract. Judy and Motta developed a customizable electrodeposition process for fabrication of very small metal structures on a substrate. In this process, layers of metal of various shapes are placed on the substrate, then the substrate is inserted in an electroplating solution. Some of the metal layers have power applied to them, while the rest of the metal layers are not connected to the power initially. Metal ions in the plating solution start depositing on the powered layers and a surface grows from the powered layers. As the surface grows, it will touch metal layers that were initially unpowered, causing them to become powered and to start growing with the rest of the surface. The metal layers on the substrate are known as seed layer patterns, and different seed layer patterns can produce different shapes. This paper presents a mathematical model, a forward simulation method, and an inverse problem solution for the growth of a surface from a seed layer pattern. The model describes the surface evolution as uniform growth in the direction normal to the surface. This growth is simulated in two and three dimensions using the level set method. The inverse problem is to design a seed layer pattern that produces a desired surface shape. Some surface shapes are not attainable by any seed layer pattern. For smooth attainable shapes, we present a computational method that solves this inverse problem.

Key words. electrodeposition, design, simulation, level set method

AMS subject classifications. 34A65, 35R35

DOI. 10.1137/070706501

1. Electrodeposition.

Judy and Motta [7] developed a customizable electrodeposition process for fabrication of small metal structures on a substrate. As pictured schematically in Figure 1.1, two plates of metal (nickel) are placed in an electroplating solution, and voltage is applied across the two plates. This causes Ni ions to separate from the anode and deposit on the cathode. By insulating areas of metal on the cathode, they can control where the metal gets deposited. Altering the exposed metal allows them to produce different shapes. The patterns of exposed metal are called seed layer patterns. For complicated seed layer patterns, it may be difficult to predict the final shape of the object. Thus the experimental procedure may require some trial and error to determine the seed pattern that attains a desired shape. One application of this process is the fabrication of neural probes used to stimulate the brain in Parkinson's disease research [6]. When used in preclinical experiments involving small animal models (e.g., rat or mouse), the probes should be very small (e.g., a few 100 μm) so they can be inserted into the brain with precision and minimal

[*] Received by the editors October 26, 2007; accepted for publication (in revised form) March 19, 2008; published electronically January 28, 2009.

http://www.siam.org/journals/siap/69-4/70650.html

[†] Level Set Systems, Los Angeles, CA 90272 (pthiyana@gmail.com).

[‡] Department of Mathematics, University of California, Los Angeles, CA 90095-1555 (caflisch@math.ucla.edu). The research of this author was supported by ARO through grant DAAD19-02-1-0336 and by the National Science Foundation (NSF) through grant DMS-0707557.

[§] Innovative Micro Technology, Goleta, CA 93117 (pmotta@ieee.org).

[¶] Electrical Engineering Department, University of California, Los Angeles, CA 90095-1594 (jack.judy@ucla.edu). The research of this author was supported in part by the NSF-supported UCLA Neuroengineering IGERT (grant 9972802) and the National Institutes of Health (NIH)-funded UCLA Morris K. Udall Center of Excellence for Parkinson's Disease Research (grant NS38367).

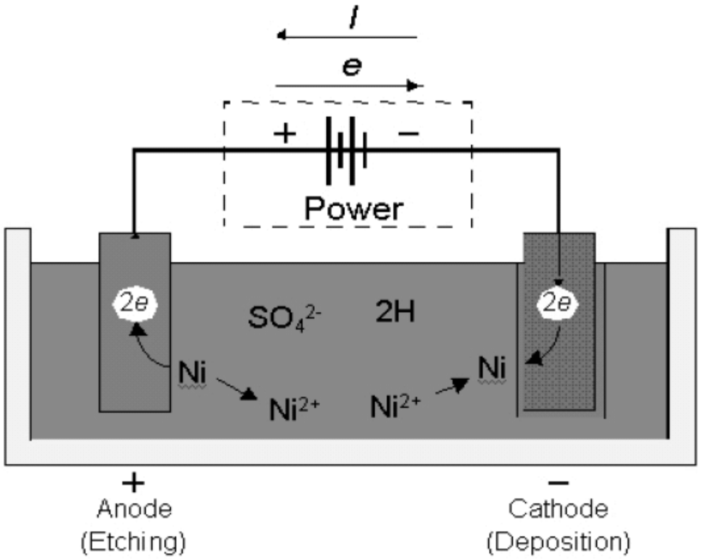


FIG. 1.1. *Experimental apparatus for plating a metal surface.*

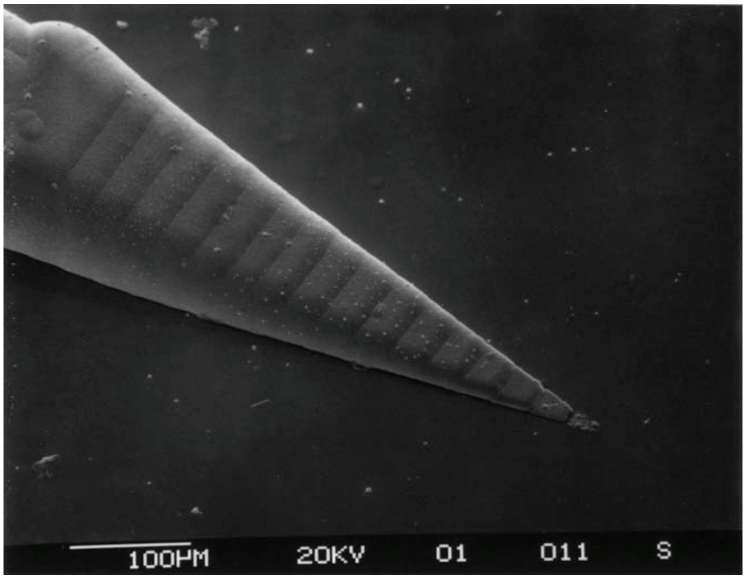


FIG. 1.2. *Experimentally formed micromachined probe shaft using a customizable electroplating process* [7].

damage [7]. Figure 1.2 shows an example of a needle-shaped object produced in [7].

A mathematical model of electrodeposition may simplify and accelerate the design of seed layer patterns and reduce the number of experimental trials needed to attain a desired shape. Our model will employ the simplifying assumption that the surface grows uniformly in the normal direction. We will not consider nonuniform growth or possible diffusion of the growing surface. Experimental parameters, such as

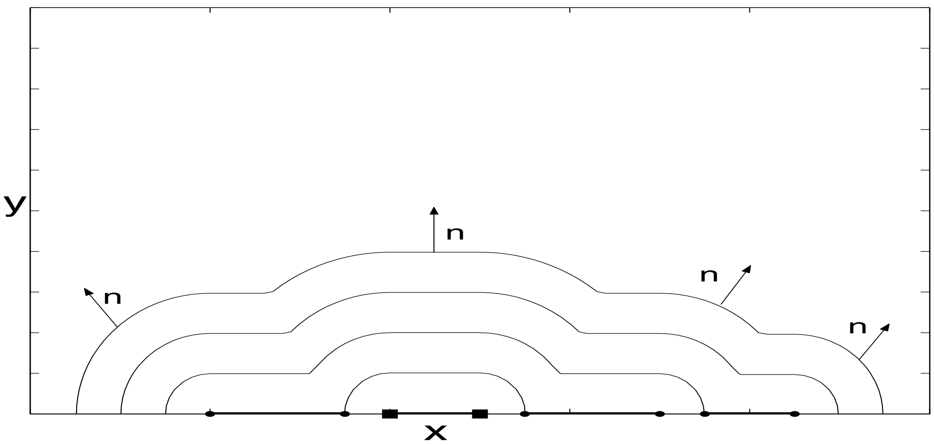


FIG. 1.3. *Stages of growth of the seed layer pattern. The segment with squares on the ends is powered and the segments with circles on ends are unpowered.*

temperature, will be also be ignored. These simplifications allow for a solution which is computationally fast and enable us to solve the inverse problem of determining a seed layer pattern that will produce a given shape.

Figure 1.3 shows a two-dimensional (2D) cross section of an object at various times as it grows uniformly in the normal direction, and indicates how the seed layer pattern controls the final shape. Initially, the segment with squares on the ends is a powered metal layer and the segments with circles on the ends are metal layers that are isolated from the power source. Although the object grows from the powered segment, it eventually touches an unpowered segment, which then becomes powered and starts to grow with the surface. As each unpowered segment is contacted by the growing object, it becomes powered and starts to grow.

The remainder of this paper is organized as follows: Section 2 describes the level set method that is the principal computational technique of our simulation method. Sections 3 and 4 describe the simulation method in two and three dimensions, respectively. The inverse design method in two and three dimensions is presented in sections 5 and 6, respectively. The methods for two and three dimensions are presented separately, since the 2D method is considerably simpler than the three-dimensional (3D) method. Conclusions and further work are discussed in section 7.

2. Level set method. The level set method is a way to represent complex geometric interfaces or shapes that evolve in time, with several nice properties, including the capability to easily merge shapes. The idea behind the level set method is to represent the interface (boundary) $\Gamma(t)$ of the object $\Omega(t)$ implicitly as the zero level set of a function ϕ, usually referred to as a level set function [8] (i.e., $\Gamma(t) = \{\mathbf{x} : \phi(\mathbf{x}, t) = 0\}$). We also set $\phi(\mathbf{x}, t) < 0$ inside $\Omega(t)$, and $\phi(\mathbf{x}, t) > 0$ outside $\Omega(t)$. This representation of $\Gamma(t)$ allows complicated topological changes, provides information about the interface, and makes computation of the evolving interface straightforward. For velocity $\mathbf{v}$ of the boundary $\Gamma(t)$, the level set function ϕ describes the evolution of $\Gamma(t)$ if ϕ satisfies

$$\phi_t + \mathbf{v} \cdot \nabla\phi = 0, \tag{2.1}$$

$$\{\mathbf{x} : \phi(\mathbf{x}, 0) = 0\} = \Gamma(0), \tag{2.2}$$

with the additional condition that ϕ is negative inside $\Omega(0)$ and positive outside $\Omega(0)$.

In the case of electrodeposition, the object $\Omega(t)$ grows normal to itself at a constant speed (i.e., the velocity is $\mathbf{v} = \nu\mathbf{n}$, in which ν is a constant and $\mathbf{n}$ is the outward unit normal vector to Γ). By rescaling time, we may take $\nu = 1$. Since the gradient $\nabla\phi$ of ϕ is in the outward normal direction $\mathbf{n}$, then $\mathbf{v} \cdot \nabla\phi = |\nabla\phi|$ and the PDE for ϕ becomes

$$\phi_t + |\nabla\phi| = 0. \tag{2.3}$$

A solution of (2.3) satisfying (2.2) is the function

$$\phi(\vec{x}, t) = d(\vec{x}) - t, \tag{2.4}$$

where $d(\vec{x})$ is the signed distance function for the initial object $\Omega_0 = \Omega(0)$ with boundary Γ_0. The function $d(\vec{x})$ is defined as

$$d(\vec{x}) = \begin{cases} -\min_{\vec{s}\in\Gamma_0} |\vec{x} - \vec{s}|\,, & x \in \Omega_0, \\ 0, & x \in \Gamma_0, \\ \min_{\vec{s}\in\Gamma_0} |\vec{x} - \vec{s}|\,, & x \notin \Omega_0, \end{cases} \tag{2.5}$$

and satisfies $|\nabla d| = 1$. It follows that $|\nabla\phi| = 1$ and $\phi_t = -1$ and that the level set $\phi = 0$ is the boundary Γ_0, so that ϕ solves (2.3).

For this model, it will be necessary to take the union of two objects and merge their boundaries. This can be done as follows: Let ϕ_1 and ϕ_2 be level set functions representing the interfaces $\Gamma_1 = \partial\Omega_1$ and $\Gamma_2 = \partial\Omega_2$, respectively. Then $\phi = \min(\phi_1, \phi_2)$ is a level set function representing the merged interface $\Gamma = \partial\Omega$ of the combined object $\Omega = \Omega_1 \cup \Omega_2$. It is important to note that merging two signed distance functions d_1 and d_2 will produce a level set function that is not necessarily a signed distance function (i.e., if $\phi = \min(d_1, d_2)$); then it is possible that $|\nabla\phi| \neq 1$ at some points inside the merged object. However, $|\nabla\phi(\vec{x})| = 1$ for $\vec{x}$ in the exterior of the interface (i.e., where $\phi(\vec{x}) > 0$). Since the interface is moving in the outer normal direction, the interface motion is still correctly described by the equation $\phi_t + 1 = 0$.

Previous applications of the level set method are found in [5, 9] for electrodeposition and in [1, 2, 3, 4] for more general deposition and materials processing.

3. Forward problem in 2D.

3.1. General idea of the solution for the forward 2D problem.

In the 2D case, the domain is $[a_1, a_2] \times [0, b_2]$. The seed layer is composed of powered and unpowered line segments placed on the x-axis, as illustrated in Figure 1.3. Suppose that there is only one powered line segment, given as $[p, q]$, and the unpowered line segments are given as $[L_i, R_i]$ for $1 \leq i \leq N$. Construct a level set function $\phi_0(x, y)$ that represents the powered line segment $[p, q]$ and solve (2.3) for $\phi(x, y, t)$, as described in section 2. The interface will grow from the powered line segment in the normal direction. After a time interval t, the interface comes in contact with (or has passed over) one of the unpowered line segments $[L_j, R_j]$ if $\phi(L_j, 0) \leq 0$ or $\phi(R_j, 0) \leq 0$. When this occurs, the jth unpowered segment becomes powered and starts to grow with the rest of the interface, which is simulated as follows:

1. Construct a level set function $\psi(x, y)$ representing the line segment $[L_j, R_j]$.
2. Merge the two interfaces by setting $\phi(x, y) = \min(\phi, \psi)$.

Now continue with the evolution starting from with the updated ϕ until another unpowered segment is contacted by the growing interface, then repeat the two steps above. For the case of multiple powered segments, steps similar to connecting an unpowered segment to the rest of the interface are performed. Suppose there are M powered segments, given by $[p_i, q_i]_{i=1}^M$ with corresponding level set functions $\phi_i(x, y)$. Then the level set function for all of the powered segments is $\phi_0(x, y) = \min_{1 \leq i \leq M}\{\phi_i(x, y)\}$, which provides the initial condition for the level set PDE (2.3).

3.2. Implementation of the solution for the forward 2D problem. In order to solve this problem computationally, space and time are discretized, with a uniform grid in space. If $[L, R]$ is a line segment, then a distance function representing it is

$$d_{[L,R]}(x, y) = \begin{cases} \sqrt{(x-L)^2 + y^2}, & x < L, \\ y, & L \leq x \leq R, \\ \sqrt{(x-R)^2 + y^2}, & x > R. \end{cases} \tag{3.1}$$

As described in section 2, the solution at discrete times $t_n = ndt$ is $\phi^{n+1} = \phi^n - dt$. The time step is chosen to be $dt = 0.1dx$, which ensures that the unpowered segments get turned on at accurate times. The initial data ϕ^0 is the distance function to the powered line segments. Contact with unpowered line segments is performed by the method of section 3.1.

Summarizing the above, if the powered segments are $[p_i, q_i]_{i=1}^M$ and the unpowered segments are $[L_i, R_i]_{i=1}^N$, the algorithm for forward growth is as follows:

```
1. Compute $\phi_0(x, y) = \min_{1 \leq i \leq M}\left(d_{[p_i,q_i]}(x, y)\right)$.
2. Set $n = 0$. Repeat the following until the final growth time $T_g$ is reached
   (i.e., when $n * dt = T_g$).
   • Set $\phi^{n+1} = \phi^n - dt$.
   • Find all unpowered segments that become powered during this time
     interval and merge them into the interface. For each unpowered segment
     $[L_k, R_k]$ do the following:
       – If $\phi^{n+1}(L_k, 0) \leq 0$ or $\phi^{n+1}(R_k, 0) \leq 0$, then compute $d_{[L_k,R_k]}(x, y)$
         and set $\phi^{n+1}(x, y) = \min\left(\phi^{n+1}(x, y), d_{[L_k,R_k]}(x, y)\right)$.
   • Set $n = n + 1$.
```

4. Forward problem in 3D.

4.1. General idea of the solution for the forward 3D problem. In the 3D case, the domain is $[a_1, a_2] \times [b_1, b_2] \times [0, c_2]$. The seed layers are on the $z = 0$ plane and can now take on any 2D shape, as illustrated in Figure 4.1. For an arbitrary shaped seed layer, initial construction of a level set function is a key step and is discussed in the section 4.3.

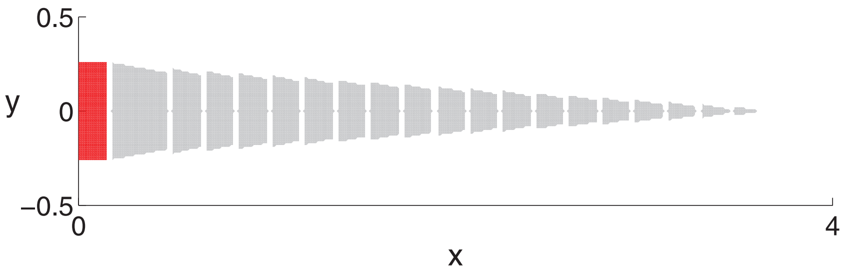


FIG. 4.1. *Seed layer used to grow the 3D needle-shaped object in Figure 1.2. The rectangle is powered and the trapezoids are the unpowered layers.*

Subsequent evolution of the level set function is essentially the same as in the 2D case: Suppose a level set function, $\phi_0(x, y, z)$, representing the powered layers is given. Using $\phi_0(x, y, z)$ as the initial condition, ϕ is evolved in time as specified by the level set PDE. The interface grows from the powered layer(s) in the normal direction. When the interface comes in contact with unpowered layers, they become powered and start to grow with the interface. In 3D, the boundaries of the layers consist of curves rather than points and some unpowered layers may have multiple boundaries, such as a ring shaped layer. Let γ be the boundary of one of the unpowered layers P. If $\min_\gamma \phi \leq 0$, then that unpowered layer has been contacted by the growing interface, it becomes powered, and it starts to grow with the interface. Merger of the layer P with the interface is performed through the following steps:

1. Construct a level set function $\psi(x, y, z)$ representing the layer P.
2. Merge the layer into the interface by setting $\phi = \min(\phi, \psi)$.

Continue with the evolution of ϕ until the growing interface contacts another unpowered layer, then repeat the two steps above.

4.2. Implementation of forward 3D method. In order to solve this problem computationally, discretize the space and time domains using a uniform grid in space. Since the seed layers lie in the plane $z = 0$, the 3D distance function $d_{3D}(x, y, z)$ is related to the 2D distance function $d_{2D}(x, y)$ within the plane $z = 0$ by

$$d_{3D}(x, y, z) = \sqrt{[\max(0, d_{2D}(x, y))]^2 + z^2}. \tag{4.1}$$

For an arbitrarily shaped seed layer U, we construct an approximate distance function $d_{2D}(x, y)$ by approximating P by an N-sided polygon with vertices $\{(x_i, y_i)\}_{i=1}^N$. A method to construct distance functions for polygons is discussed in section 4.3. Detecting contact requires evaluation of the level set function ϕ at the boundaries of the unpowered layer. In the numerical method, this is checked at a discrete set of points along the polygonal boundary. If $\phi \leq 0$ on any of the discrete points on the boundary, the unpowered polygonal layer has been contacted by the interface, becomes powered, and starts to grow with the interface. This is done by performing the following two steps:

1. Construct a distance function $d(x, y, z)$ representing the layer.
2. Take the union of the two interfaces by setting $\phi = \min(\phi, d)$.

As described in section 2 and the 2D case, the iteration to evolve the interface is $\phi^{n+1} = \phi^n - dt$, where ϕ^0 is the distance function representing the powered layer(s). The boundary points of each unpowered layer are checked for contact each time step.

Summarizing the above, let $\{P_i^p\}_{i=1}^M$ be the set of powered polygonal layers and $\{P_i^u\}_{i=1}^N$ the set of unpowered polygonal layers of a given seed layer pattern. The following steps detail forward growth in three dimensions:

1. Set $\phi^0(x, y, z) = \min_{1\leq i\leq M}(d_i(x, y, z))$ in which $d_i(x, y, z)$ is the distance function (see section 4.3) for the ith powered polygonal layer P_i^p.
2. Set $n = 0$. Repeat the following until the final growth time T_g is reached (i.e., when $n * dt = T_g$).
    - Set $\phi^{n+1} = \phi^n - dt$.
    - Check all of the unpowered polygonal layers that have not already been merged into the growing shape for contact by the interface. If $\phi^{n+1} \leq 0$ at one of the discretized boundary points of an unpowered polygonal layer, it has been contacted by the interface.
        - Suppose the kth unpowered polygonal layer P_k^u has been contacted

by the interface. Compute $d_k(x,y,z)$, and set

$$\phi^{n+1}(x,y,z) = \min\left(\phi^{n+1}(x,y,z), d_k(x,y,z)\right).$$

– Repeat the above step for any other newly contacted unpowered layer(s).

- Set $n = n + 1$.

4.3. Constructing signed distance functions for polygons. Suppose that $\{(x_k, y_k)\}_{k=1}^{N}$, $(x_k, y_k) \in [a_1, a_2] \times [b_1, b_2]$, are the vertices of a polygon traversed in either a clockwise or counterclockwise direction, with $(x_N, y_N) = (x_1, y_1)$. The goal is to create a signed distance function $d(x, y)$ whose zero level set is the polygon boundary. Since the level set computations are performed on a grid, it is only necessary to compute the distance function on the discrete set of grid points. The first step is to create a distance function $d_{k,k+1}(x, y)$ for each line segment $[(x_k, y_k), (x_{k+1}, y_{k+1})]$ (see the appendix for the formula). Note that $d_{k,k+1}(x, y) \geq 0$, since the line segment has no interior. Set $d_0(x, y) = \min_k \{d_{k,k+1}(x, y)\}$ to get a distance function $d_0(x, y)$ for the polygon defined on the grid. However, $d_0(x, y)$ is not a signed distance function since it takes positive values inside the polygon. The remaining step is to identify which grid points are inside the polygon, and negate the value of d_0 at those points.

The Jordan curve theorem states that a point is in the interior of a bounded region if the half line from that point to infinity intersects the boundary of the region an odd number of times. Conversely, the point is in the exterior of the region if the half line intersects the boundary an even number of times. The half line from the point in question to infinity can be taken in any direction. For the grid point (u, v), a natural choice is the horizontal half line $((-\infty, v), (u, v)]$. Since the polygon should be completely contained in the domain $[a_1, a_2] \times [b_1, b_2]$, only the horizontal line segment $[(a_1, v), (u, v)]$ needs to be checked for intersections with the polygon boundary. Since the boundary of the polygon is composed of line segments, it suffices to check for intersections between the horizontal line segment $[(a_1, v), (u, v)]$ and the polygon boundary segments.

In summary, the steps to form a signed distance function for a polygon on a grid are as follows:

1. Set $d_0(x, y) = \min_{1 \leq k < N}(d_{k,k+1}(x, y))$ in which $d_{k,k+1}(x, y)$ is the distance function from the line segment between vertices k and $k + 1$.
2. For each grid point (u, v), do the following:
    - Count the number of intersections of the line segment $[(a_1, v), (u, v)]$ with the polygon sides.
    - If the number of intersections is odd, set $d_0(x_i, y_j) = -d_0(x_i, y_j)$.

Once the 2D distance function $d_0(x, y)$ has been computed; the 3D distance function for the polygonal layer is $d_0(x, y, z) = \sqrt{[\max(d_0(x, y), 0)]^2 + z^2}$.

4.4. Results for forward 3D method. Figure 4.1 shows the seed layers that were used to grow the 3D needle shaped object pictured in Figure 1.2. The leftmost (rectangular) layer was powered and the others were unpowered initially. Figure 4.2 shows the results from simulation, using the method described in section 4.2, applied to the seed layer pattern given in Figure 4.1. The simulated object (Figure 4.2) is in excellent agreement with the object that was grown in the laboratory (Figure 1.2).

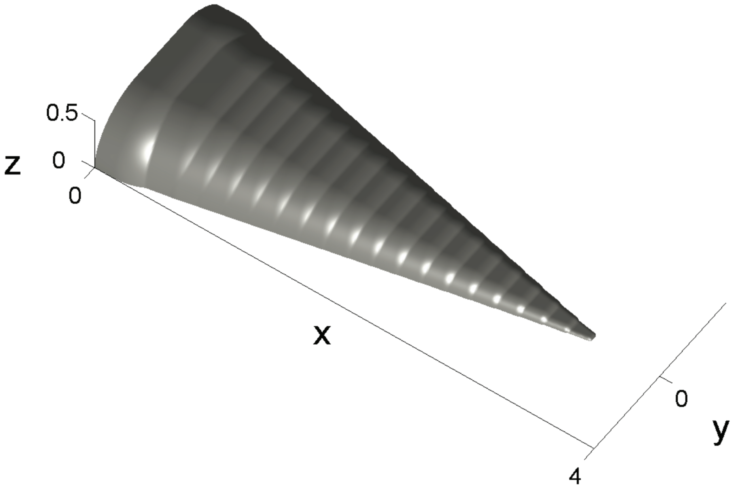


FIG. 4.2. *Simulated growth of the needle seed layer pattern of Figure* 4.1*, using the algorithm described in section* 4.2.

5. Inverse problem in two dimensions.

5.1. Problem description and solution. For the inverse problem, a shape is given, and the goal is to determine the seed layer pattern that would approximate that shape. In addition, we find that some shapes are not attainable.

Let the desired final shape Γ be given by $y = F(x)$, contained in the domain $[a_1, a_2] \times [0, b_2]$. The results of section 3 show that the ends of a line segment grow as quarter circles. The first step in solving the inverse problem is to approximate the curve Γ by an envelope of circles that are tangent to Γ and that have centers on the x-axis. Assume that $F \in C^1([a_1, a_2])$. For $s \in [a_1, a_2]$, the circle tangent to F at the point $(s, F(s))$ has center $(c_s, 0)$ and radius R_s given by the formulas

$$c_s = F'(s)F(s) + s, \tag{5.1}$$

$$R_s = F(s)\sqrt{[F'(s)]^2 + 1}. \tag{5.2}$$

A discrete representation of the curve Γ is given by points $\{(x_i, F(x_i))\}_{i=1}^N$ for a given set of points $x_i \in [a_1, a_2]$. Equations (5.1)–(5.2) then determine centers $\{(c_i, 0)\}_{i=1}^N$ and radii $\{R_i\}_{i=1}^N$ for a set of circles $\{C_i\}_{i=1}^N$ that are tangent to F at the points $(x_i, F(x_i))$. Remove any circle C_i whose center is not in the domain (i.e., $c_i \notin [a_1, a_2]$)) to get a possibly smaller sets of centers $\{(c_j, 0)\}_{j=1}^M$ and radii $\{R_j\}_{j=1}^M$, with $M \leq N$. See Figure 5.1 for an example of an envelope of circles tangent to the function $F(x) = \frac{1}{2}\sqrt{1 - x}$.

For simplicity, we assume that initially there is only a single powered point. It starts growing as a half circle whose radius is equal to the time that point has been powered (since the normal velocity is 1). More generally, the radius of a circle in the envelope is equal to the total time over which the circle should be powered. It follows that the total growth time $t = T_G$ should equal the maximal radius of the circles (i.e., $T_G = \max_j \{R_j\}$). Furthermore, since the jth circle should grow for time R_j, then its start time should be $T_j = T_G - R_j$. The circle C_j gets powered at time T_j if its

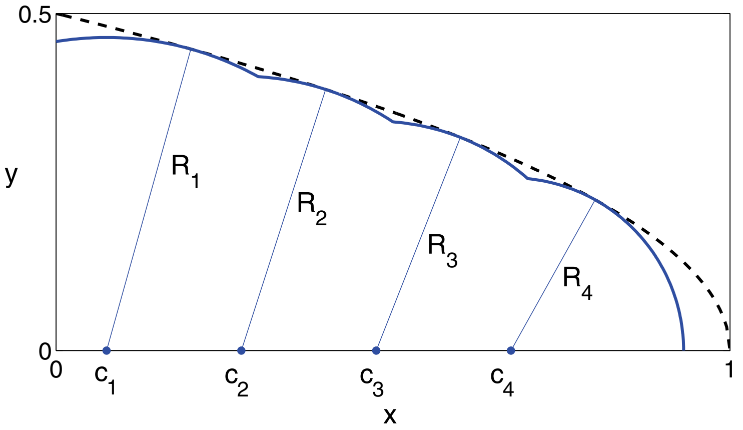


FIG. 5.1. *An envelope of* 4 *circles are used to form an approximation of* $y = F(x)$.

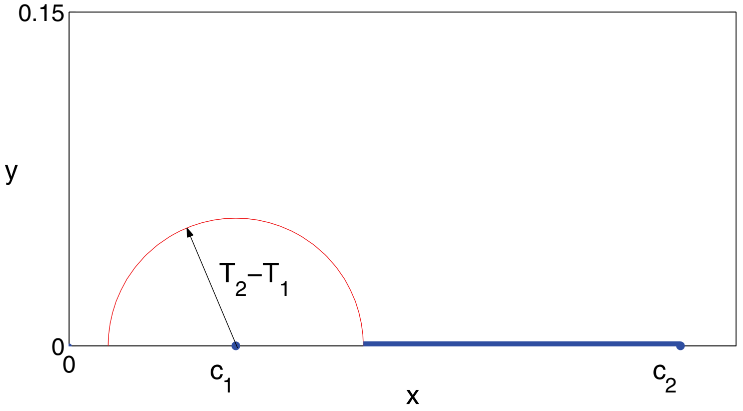


FIG. 5.2. *Shows the first circle at time* T_2 *and the unpowered segment that would power the second circle at time* T_2.

center is the endpoint of a segment that is contacted at time T_j by the circle that is growing from the center $(c_{j'}, 0)$ to its left or right that was powered before T_j.

This is illustrated in Figures 5.1–5.3. In the example shown in Figure 5.1, R_1 is the maximum radius, so $T_G = R_1$, $T_1 = 0$ and the point $(c_1, 0)$ should be powered initially. The point $(c_2, 0)$ should be powered at time T_2. Figure 5.2 shows the first circle at time T_2 turning on the point $(c_2, 0)$ by contacting the left edge of the segment $[\ell, c_2]$ at time T_2 with $\ell = c_1 + T_2 - T_1$. Figure 5.3 shows the interface at time T_3, including the growth of the unpowered segment calculated to turn on the second circle.

To summarize the procedure in general for finding the unpowered segment for the kth circle centered at $(c_k, 0)$, with start time T_k, it is necessary to find the closest circle to the kth circle with an earlier start time. Denote the closest circle as the pth circle with center $(c_p, 0)$ and start time T_p. If $c_p < c_k$, then the unpowered segment is on the left side of c_k and is given by $[\ell, c_k]$, where $\ell = c_p + T_k - T_p$. If $c_p > c_k$, then the unpowered segment is on the right side of c_k and is given by $[c_k, r]$, where $r = c_p - (T_k - T_p)$.

Three sufficient conditions on the curve $\Gamma = \{y = F(x)\}$ must be satisfied to

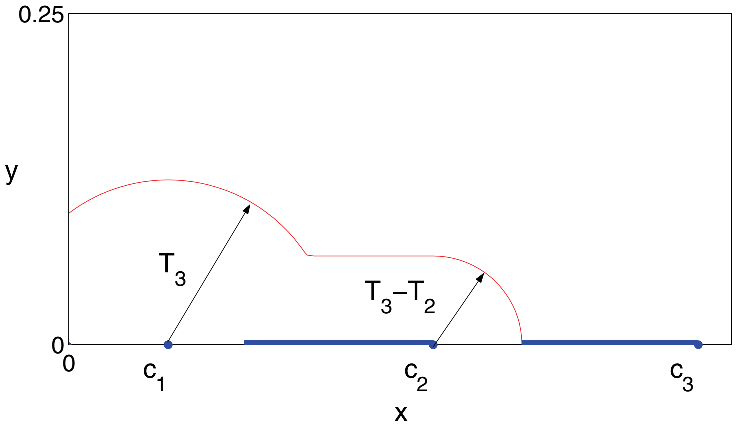


FIG. 5.3. *Shows the interface at time T_3, and the unpowered segment that powers the third circle at time T_3.*

enable a solution to the inverse problem:

1. $F \in C^1([a_1, a_2])$.
2. Γ is the envelope of circles with centers on the x-axis. Each circle intersects Γ at tangent points from below (i.e., the circle lies below Γ).
3. F can have only one local maxima, or if it has multiple local maxima, they must all have the same value.

Although there may exist a seed layer pattern for a given shape that is only in $C([a_1, a_2])$ (but still satisfies the other two conditions), the given algorithm requires that the derivative be continuous as well. The second condition is a necessary one. If that condition is violated, then the shape given by $F(x)$ will not be attainable. Figure 5.4 shows the function $F(x) = \frac{1}{4}\sin(\pi x^2)$, and a circle tangent to it at $(0.8, F(0.8))$ that also crosses F at another point, making that shape unattainable. The present model allows for powered points to be powered by an external source only at the same time. For example, there cannot be one point x_1 powered at time $t = 0$ and another point x_2 independently powered (meaning not powered by contact of the growing surface) at some later time. The third condition is therefore necessary since it ensures that an approximation of the given shape can be attained without needing powered points that start at different times.

5.2. Justification for the construction of the unpowered segments. In constructing the unpowered segments, it was assumed that the point $\ell = c_p + T_k - T_p$ was in between the two circle centers c_p and c_k (this assumes that the contact point is to the left of c_k, but the proof is similar if it were to the right). Here it will be shown that this is always the case if the attainability conditions of section 5.1 are satisfied. Showing that $\ell \in [c_p, c_k]$ is the same as showing that $\ell = c_p + T_k - T_p \le c_k$, or

$$T_k - T_p \le c_k - c_p.$$

Suppose this were not true, so that $T_k - T_p > c_k - c_p$. Using the relationships $T_k = T_g - R_k$ and $T_p = T_g - R_p$ gives $T_k - T_p = R_p - R_k$. Then $R_p - R_k > c_k - c_p$, or

$$R_p > d + R_k,$$

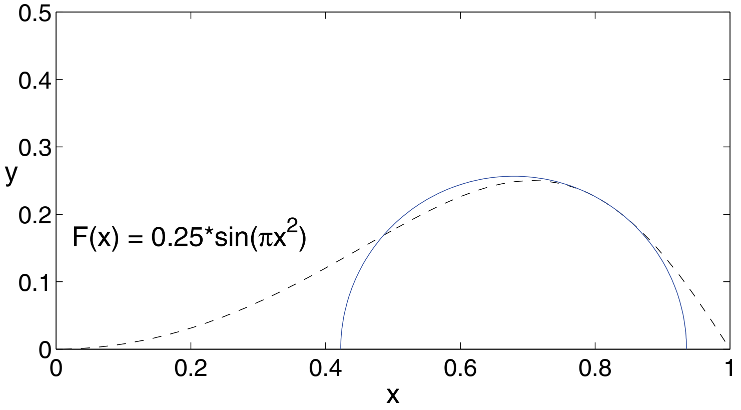


Fig. 5.4. *Example of a shape that is not attainable. The circle is tangent to F at* $(0.8, F(0.8))$*, but it crosses F at another point. This function is not attainable using the given procedure for calculating the seed layer pattern.*

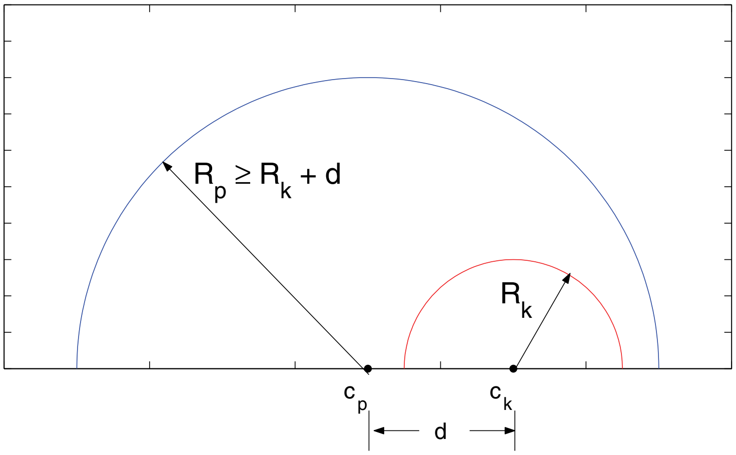


Fig. 5.5. *The figure shows that the circle centered at c_p completely encloses the circle centered at c_k if $T_k - T_p \geq c_k - c_p$.*

where $d = c_k - c_p$ is the distance between the two circle centers. This means that the circle centered at c_p has a radius so large that it completely encloses the circle centered at c_k (see Figure 5.5). This is impossible, since both circles are tangent to F from below by the second attainability condition in section 5.1. This contradiction shows that $\ell = c_p + T_k - T_p \leq c_k$.

5.3. Results of inverse 2D problem method. Figure 5.6 shows the solution of the inverse problem for the function $F(x) = x(x-1)^2$. For solution of the inverse problem, a single powered point and eight unpowered segments are allowed. The figure shows the desired function (dashed line), the calculated seed layer pattern, and the resulting growth from the seed layer pattern. The result is in good agreement with the target curve. Better agreement could be achieved by allowing more segments in the seed layer pattern.

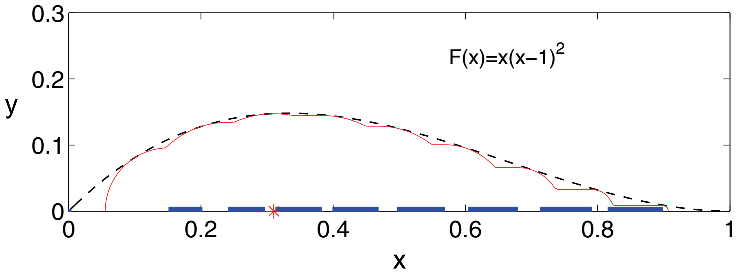


FIG. 5.6. *The desired shape is the dotted curve. The seed layer was calculated using the inverse procedure, and the growth of the seed layer is shown by the solid curve. The star is the powered point, and the line segments are the unpowered segments. Nine circles were used for the approximation.*

6. Inverse problem in three dimensions.

6.1. A continuous approach. The desired shape Γ is defined by the function $z = F(x, y)$ for (x, y) in the substrate domain $[a_1, a_2] \times [b_1, b_2]$. The goal is to find a seed layer pattern of polygonal layers that will grow to form an approximation of Γ. Instead of a discrete seed layer, consider a continuous seed layer, in which each point in the substrate can be powered at any time. Suppose that this continuous seed layer can grow a shape that exactly matches the given shape Γ at time T_g. This continuous seed layer can then be described by a function $T(x, y)$, which is defined to be the time at which power should be applied to the point (x, y). Although a continuous seed layer may be impractical, it is a useful theoretical construct.

Let $t_j \in (0, T_g)$ for $j = 1, \ldots, n$ be an increasing set of times such that $0 = t_0$, and define the contour curves

$$\gamma_i = \{(x, y) \in [a_1, a_2] \times [b_1, b_2] : T(x, y) = t_i\}\,. \tag{6.1}$$

Also, define β_i to be the curve that results from growing γ_i in the outward normal direction (i.e., the direction of increasing T) for time $(t_{i+1} - t_i)$.

The initial powered "layer" S_0 is defined to be the curve (or point) γ_0. For $m > 0$, the seed layer S_m is then defined to be the region in between the two curves γ_m and β_{m-1}. With this definition, we find that S_m is contacted by the growing surface and turns on at time t_m as desired. By its definition, S_0 is powered at time t_0. The rest follows by iteration: Since the curve γ_{m-1} becomes powered at time t_{m-1} it grows outward at normal speed 1, and at time t_m it hits the curve β_{m-1} which turns on S_m, which is the region in between β_{m-1} and γ_m. This is illustrated in Figure 6.1, which shows two contours γ_m and γ_{m-1} of some function $T(x, y)$. Also shown is β_{m-1}, which is the resulting curve of growing γ_{m-1} in the normal direction for time $t_m - t_{m-1}$.

All of the unpowered layers are constructed in the same way. When the above method is discretized, the powered curve γ_0 will usually not be a curve, but individual points. From a fabrication viewpoint, both individual points and curves as powered layers are not desirable, and it is better to have an actual polygonal powered layer. One simple approach to do this will be discussed later.

6.2. Discretizing and implementing the continuous approach. There are several tasks involved in solving the inverse problem computationally. Each of the following sections describes a task and a method to complete the task. Then, in section 6.2.7, the methods are brought together to give the solution to the inverse problem.

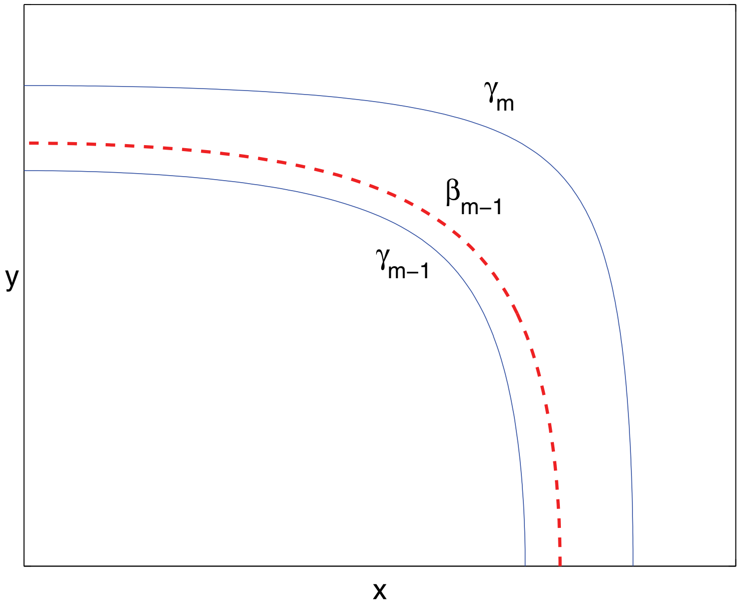


FIG. 6.1. *Two contours γ_{m-1} and γ_m of $T(x,y)$ are shown as solid curves. β_{m-1} is shown as the dashed curve. The unpowered layer is the region in between β_{m-1} and γ_m.*

6.2.1. Relationship between $T(x,y)$ and $F(x,y)$. A key component of the previous section was knowing $T(x,y)$ for a given shape Γ defined by $z = F(x,y)$. An equivalent condition will be stated and used here. Consider a seed layer that is only a single powered point. Since the growth is uniform in the normal direction with speed 1, the powered point will grow as a hemisphere, with a radius equal to the time the point has been powered. Let the given shape $z = F(x,y)$ satisfy the following conditions:

1. $F \in C^1([a_1,a_2] \times [b_1,b_2])$.
2. Γ can be formed as an envelope of hemispheres (with centers on the $z = 0$ plane) that are tangent to Γ from below (i.e., the sphere lies below the surface $z = F(x,y)$).
3. F has only one local maximum, or if it has multiple local maxima, F has the same value at all local maxima.

Under the above conditions, there should exist a function $R(x,y)$, $(x,y) \in [a_1,a_2]\times[b_1,b_2]$, that gives the radius of the sphere with center (x,y) that is tangent to the given shape F. A continuous seed layer can be defined by $R(x,y)$, since $R(x,y)$ is the length of time that a point (x,y) should be powered. Define $T_g = \max_{(x,y)} \{R(x,y)\}$, which is the length of time that power should be applied to the seed layer. Then $T(x,y) = T_g - R(x,y)$ is the time at which the point (x,y) should become powered. Therefore, finding $T(x,y)$ is equivalent to finding $R(x,y)$. By considering the equation of a sphere with center (x,y) on the $z = 0$ plane, the relationship between $R(x,y)$ and F is given by the following three equations:

$$x = F_x(x_s,y_s)F(x_s,y_s) + x_s, \tag{6.2}$$

$$y = F_y(x_s,y_s)F(x_s,y_s) + y_s, \tag{6.3}$$

$$R(x,y) = \sqrt{(x-x_s)^2 + (y-y_s)^2 + [F(x_s,y_s)]^2}. \tag{6.4}$$

The point $(x_s, y_s, F(x_s, y_s)$ is the point of tangency of the sphere to the surface $z = F(x, y)$.

6.2.2. Computing $\boldsymbol{T(x,y)}$. The first step in numerically solving the inverse problem is to discretize the domain of the substrate $[a_1, a_2] \times [b_1, b_2]$. A uniform grid is used. Let $N_x + 1$ be the number of grid points in x and $N_y + 1$ the number of grid points in y. Then the grid sizes are $dx = (a_2 - a_1)/N_x$ and $dy = (b_2 - b_1)/N_y$, and the grid points are $\{(x_i, y_j) = (a_1 + i * dx, b_1 + j * dy)\}$ for $0 \leq i \leq N_x$ and $0 \leq j \leq N_y$. The radius value $R(x_i, y_j)$ on the grid point (x_i, y_j) is found by solving the nonlinear equations (6.2)–(6.3). Numerical experiments indicate that the following simple functional iteration converges fairly quickly:

$$x_s^{n+1} = x_i - F_x(x_s^n, y_s^n)F(x_s^n, y_s^n), \tag{6.5}$$

$$y_s^{n+1} = y_j - F_y(x_s^n, y_s^n)F(x_s^n, y_s^n). \tag{6.6}$$

A good initial guess is $x_s^0 = x_i$ and $y_s^0 = y_j$: Once (x_s, y_s) is found, (6.4) is used to compute $R(x_i, y_j)$. This process is repeated to compute R for each grid point in the discretized domain. Then the growth time is determined by $T_g = \max_{[0 \leq i \leq N_x, 0 \leq j \leq N_y]} \{R(x_i, y_j)\}$ and the function T is given by $T(x_i, y_j) = T_g - R(x_i, y_j)$ on the discretized domain.

6.2.3. Calculating contour curves. Although any conventional methods can be employed to calculate the contour curves of $T(x, y)$, a simple method is presented here. Let $T(x, y)$ be given on the grid points defined in section 6.2.2. Linear interpolation is used to find the points of a contour curve $\gamma = \{(x, y) : T(x, y) = \alpha\}$ on a triangulated grid. Consider the lower triangular cell with corners (x_i, y_j), (x_{i+1}, y_j), and (x_{i+1}, y_{j+1}). If γ passes through that lower triangular grid cell, then γ can be approximated by a line segment $L_{i,j}^{\ell}$ inside the lower triangular grid cell. Linear interpolation can determine the endpoints of the line segment $L_{i,j}^{\ell}$, which could be on the horizontal grid cell boundary $[(x_i, y_j), (x_{i+1}, y_j)]$, the vertical grid cell boundary $[(x_{i+1}, y_j), (x_{i+1}, y_{j+1})]$, or the diagonal grid cell boundary $[(x_i, y_j), (x_{i+1}, y_{j+1})]$. Every lower triangular grid cell is checked to see if γ passes through it, and if so a line segment $L_{i,j}^{\ell}$ is computed. Similarly, the upper triangular grid cells with corners (x_i, y_j), (x_i, y_{j+1}), and (x_{i+1}, y_{j+1}) are also checked to see if they contain part of γ, and if so, a line segment $L_{i,j}^{u}$ is computed. The collection of all line segments $L_{i,j}^{\ell}$ and $L_{i,j}^{u}$ is then a representation of the contour curve γ. Connectivity of this curve is guaranteed, since the end of one line segment corresponds to the beginning of another line segment. Figure 6.2 illustrates the triangulated grid and the points in the triangular grid cell boundaries that define the ends of the line segments.

6.2.4. Evolving a curve in the normal direction. One step in section 6.1 for finding an unpowered layer requires growing a contour curve of $T(x, y)$ in the outward normal direction, which is performed using the level set method. Let γ be the contour $T(x, y) = \alpha$ that is to be moved in the outward normal direction for time τ. Using the method discussed in section 6.2.3, compute a set of line segments $\{L_k\}$ that represents the contour curve γ. Following steps similar to those outlined the section 4.3, a distance function representing the polygonal approximation of γ is computed as $\psi(x, y) = \min_k \{d_k(x, y)\}$, where $d_k(x, y)$ is the distance function for the line segment L_k (a formula is given in the appendix). Note that ψ will not be a signed distance function, since $d_k(x, y) \geq 0$. In general, γ will segment the domain into two parts: the interior region (where $T(x, y) < \alpha$) and the exterior region (where

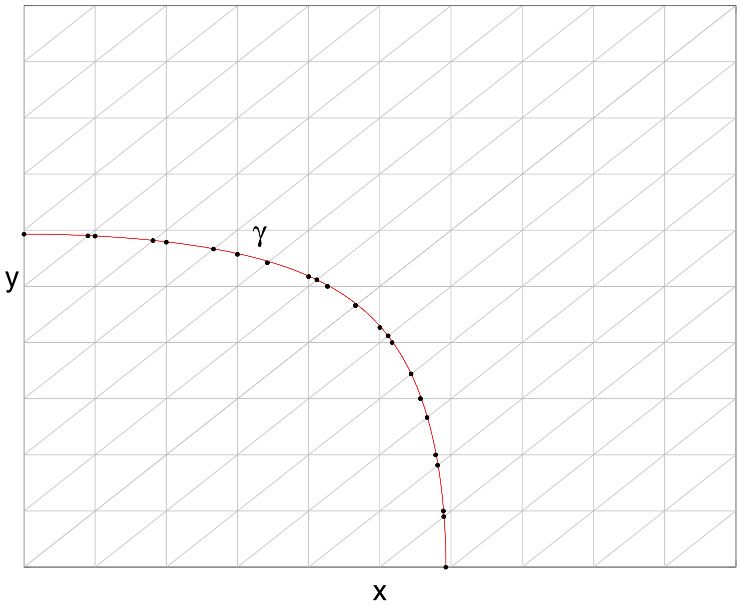


FIG. 6.2. *The circles are the endpoints of the line segments used for the discretization of the curve* γ.

$T(x, y) > \alpha$). To make a signed distance function, the value of ψ in the interior region must be negated. Therefore, a signed distance function ϕ for the curve γ is defined as

$$\phi(x, y) = \begin{cases} -\psi(x, y) & \text{if } T(x, y) < \alpha, \\ \psi(x, y) & \text{otherwise.} \end{cases}$$

Then the curve from growing γ in the normal direction for time τ is the zero level set of $\phi(x, y, 0) - \tau$. Again, the methods in section 6.2.3 can be used to form a polygonal approximation of the new curve.

6.2.5. Forming an unpowered layer from $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The final step in the construction of a layer S_m is to connect the two curves, β_{m-1} and γ_m, that define it in the $z = 0$ plane. Let $\{L_k^\beta\}$ and $\{L_k^\gamma\}$ be the set of line segments representing the curves β_{m-1} and γ_m, respectively. The unpowered layer is the region between these two curves consisting of points (x, y) with $t_{m-1} < T(x, y) < t_m$. In the simplest case, β_{m-1} and γ_m are each single connected curves. If they are closed curves, then S_m is the region between them. If they are not closed, then each of them has two endpoints that should lie on the boundary of the domain. In this case, one can easily identify the matching endpoints (by the condition $t_{m-1} < T(x, y) < t_m$) and form a curve connecting them along the domain boundary. In general, γ and β may have multiple disjoint pieces that define multiple layers (in such a case, these layers would all have the same start time). A more sophisticated approach to forming layers may have to be used to connect complicated curves, but this has not proved to be necessary in the numerous examples we have simulated.

6.2.6. The powered "layer." In the presentation above, the initially powered layer consists of a single point or curve that is discretized to a set of points. The powered points should have start time $t = 0$, so they are the points on the grid

$\{(x_k^P, y_k^P)\}_{k=1}^{N_P}$ that satisfy $T(x_k^P, y_k^P) = 0$. In order to calculate the first unpowered layer, the powered points will have to be grown for time t_1. The signed distance function for a point is (u, v) is $d_{(u,v)}(x, y) = \sqrt{(x-u)^2 + (y-v)^2}$. To grow the powered points for time t_1, the distance function for the powered points must first be computed as

$$d(x, y) = \min_{k=1,\dots,N_P} \{d_{(x_k^P, y_k^P)}(x, y)\}. \tag{6.7}$$

Then, the zero level set of $d(x, y) - t_1$ is the curve resulting from growing the powered points for time t_1.

6.2.7. Constructing the seed layer pattern. This section brings together all of the methods in the previous sections to formulate in detail a solution to the inverse problem. Let the substrate domain be $[a_1, a_2] \times [b_1, b_2]$, and assume that the surface $z = F(x, y)$ satisfies conditions (1)–(3) in section 6.2.1.

- Discretize the domain using a uniform grid, where $N_x + 1$ is the number of grid points in x and $N_y + 1$ is the number of grid points in y.
- Compute $R(x, y)$ on the grid using the method described in section 6.2.2. Define $T_g = \max_{[0 \le i \le N_x, 0 \le j \le N_y]} \{R(x_i, y_j)\}$.
- Set $T(x, y) = T_g - R(x, y)$ on the grid.
- Define a positive integer n, and choose a set of times $\{t_i\}_{i=1}^n$ such that $0 = t_0 < t_1 < t_2 < \cdots < t_n < T_g$.
- Find the powered points of the seed layer $\{(x_k^P, y_k^P)\}_{k=1}^{N_P}$ that satisfy $T(x_k^P, y_k^P) = 0$.
- Using the technique given in section 6.2.6, compute the distance function $d(x, y)$ for the powered points.
- Compute the parameterization of $\beta_0 = \{(x, y) \colon d(x, y) = t_1\}$ using the method in section 6.2.3.
- Compute the parameterization of $\gamma_1 = \{(x, y) \colon T(x, y) = t_1\}$ using the method in section 6.2.3.
- Use the method in section 6.2.5 to form the first unpowered layer that is in between β_0 and γ_1.
- For $m = 2, \dots, n$, repeat the following:
  - Grow γ_{m-1} in the outward normal direction for time $(t_m - t_{m-1})$ using the method in section 6.2.4 and label the resulting curve β_{m-1}.
  - Compute the parameterization of $\gamma_m = \{(x, y) : T(x, y) = t_m\}$ using the method in section 6.2.3.
  - Use the method in section 6.2.5 to construct the mth unpowered layer that is in between β_{m-1} and γ_m.

6.2.8. Justification for the construction of the unpowered layers. In the construction of the unpowered layers, it was assumed that the curve β_{m-1} is between γ_{m-1} and γ_m (i.e., that when γ_{m-1} moves in the normal direction for time $t_m - t_{m-1}$, it does not cross γ_m). We now prove this, assuming that the shape $z = F(x, y)$ satisfies the attainable conditions in section 6.2.1.

From any point (x^{m-1}, y^{m-1}) on γ_{m-1}, consider the line segment extending in the outer normal direction to a point (x^m, y^m) on the curve γ_m, and define the distance $d = |(x^{m-1}, y^{m-1}), (x^m, y^m)|$. The map from (x^{m-1}, y^{m-1}) to (x^m, y^m) is one-to-one if the curves are smooth and the distance between them is sufficiently small. Since the curve β_{m-1} is defined by moving γ_{m-1} a distance $t_m - t_{m-1}$, it suffices to show that $t_m - t_{m-1} \le d$.

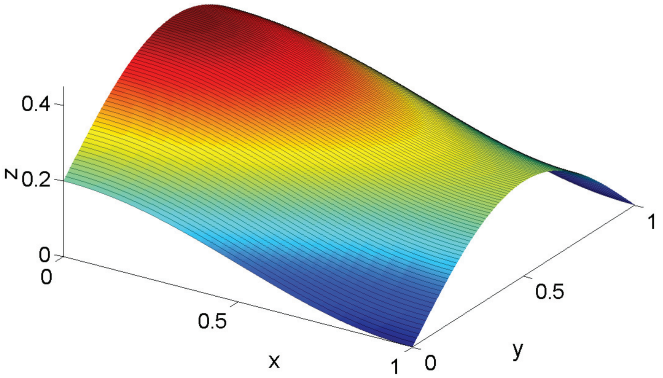


FIG. 6.3. *The desired shape is given by $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$. The color map is a function of height and is used only to aid visualization of the surface.*

Suppose that the opposite is true, so that $t_m - t_{m-1} > d$. By condition 2 in section 6.2.1, the sphere S_{m-1} centered at (x^{m-1}, y^{m-1}) with radius $R^{m-1} = T_g - t_{m-1}$ and the sphere S_m at (x^m, y^m) with radius $R^m = T_g - t_m$ are both tangent to the surface $z = F(x, y)$. It follows that $R^{m-1} \geq R^m + d$, which says that S_{m-1} strictly contains S_m, which is not possible if both are tangent from below by condition 2 in section 6.2.1. This contradiction shows that $t_m - t_{m-1} \leq d$, so that the curve β_{m-1} is between γ_{m-1} and γ_m.

6.2.9. A powered polygonal layer. The algorithm outlined in section 6.2.7 constructs the powered layer as points rather than the preferred polygonal shape. A simple way to construct a powered polygonal layer is to set the first unpowered layer as powered. The final growth time would then have to be reduced by t_1, the start time of the first unpowered layer. Doing this will increase the error of the final shape, but this error can be reduced by choosing t_1 to be small.

6.2.10. Results. Figures 6.3–6.5 show a prescribed shape $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$, the seed layer pattern calculated using the steps in section 6.2.7, and the result of the forward growth of the calculated seed layer pattern, respectively. Figures 6.6–6.8 show similar results for the prescribed shape $F(x, y) = -(x - 0.5)^2 - (y - 0.5)^2 + 0.25$. Both examples were computed on the domain $[0, 1] \times [0, 1]$ with $N_x = N_y = 100$ and $n = 12$. The computational time to construct the seed layer patterns for the above examples was about 5 seconds.

6.3. Error analysis. In this section, a computational analysis of the accuracy of the inverse algorithm of section 6.2 is presented using the two test problems in section 6.2.10 for various values of the number of grid points (N_x and N_y) and the parameter n, which determines the number of unpowered layers ($\geq n$). In order to resolve the seed layers, the number of grid points should be larger than the number of layers.

The accuracy of the inverse solution is measured by the error in the subsequent forward growth. The forward problem is performed on the computational domain $[0, 1] \times [0, 1] \times [0, 0.5]$, with a fixed number of grid points $N_x^F = 100$, $N_y^F = 100$, and $N_z^F = 50$, with grid sizes $dx^F = \frac{1}{N_x^F+1}$, $dy^F = \frac{1}{N_y^F+1}$, and $dz^F = \frac{0.5}{N_z^F+1}$. Most of the error in the forward growth itself is removed, since the start times of the layers

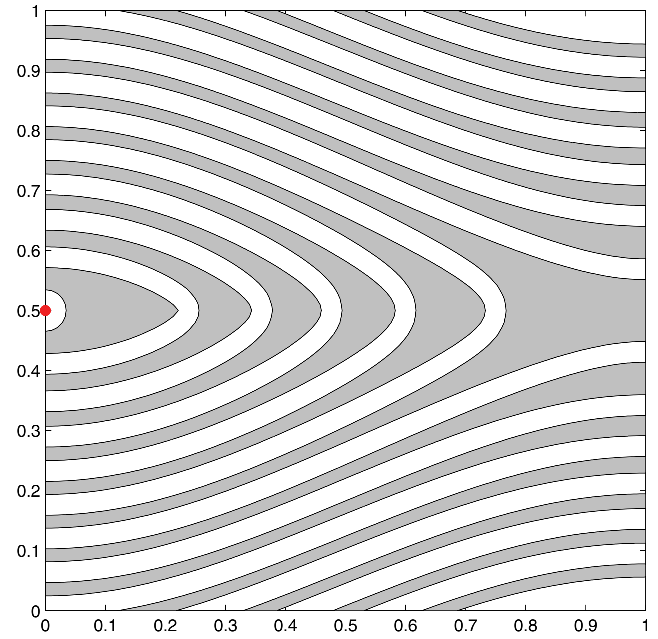


Fig. 6.4. *The calculated polygonal seed layer pattern for the shape given by* $F(x,y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$*, shown in Figure* 6.3. *The dot is the powered point.*

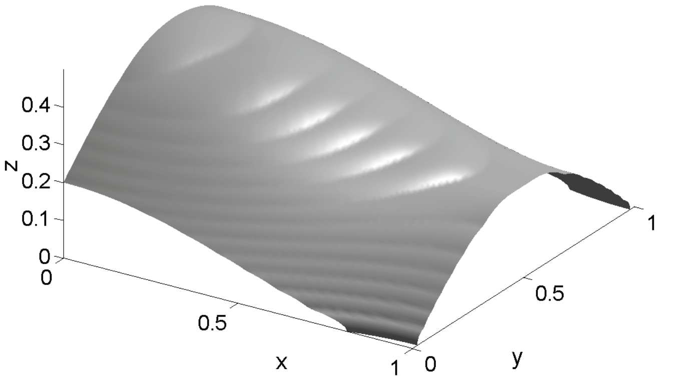


Fig. 6.5. *The forward growth of the calculated polygonal seed layer pattern in Figure* 6.4. *It compares well to the desired shape in Figure* 6.3.

are all known. The resulting level set function $\phi(x, y, z)$ is converted to a function $z = F^{\phi}(x, y)$ defined on the fixed forward grid using linear interpolation. A numerical L^2 error between F^{ϕ} and F is computed as

$$\left\|F - F^{\phi}\right\| = \sqrt{\sum_{i=0,j=0}^{N_x^F, N_y^F} \left|F(x_i, y_j) - F^{\phi}(x_i, y_j)\right|^2 dx^F dy^F}. \tag{6.8}$$

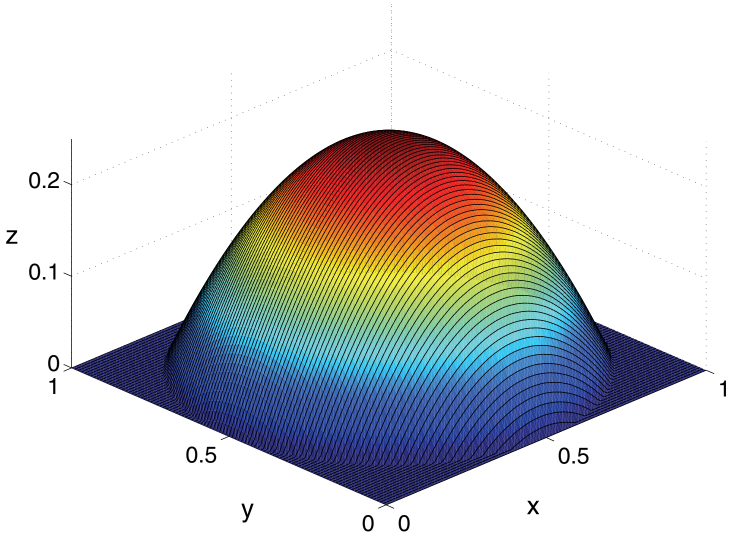


Fig. 6.6. *The desired shape is given by $F(x,y) = -(x-0.5)^2 - (y-0.5)^2 + 0.25$. The color map is a function of height and is used only to aid visualization of the surface.*

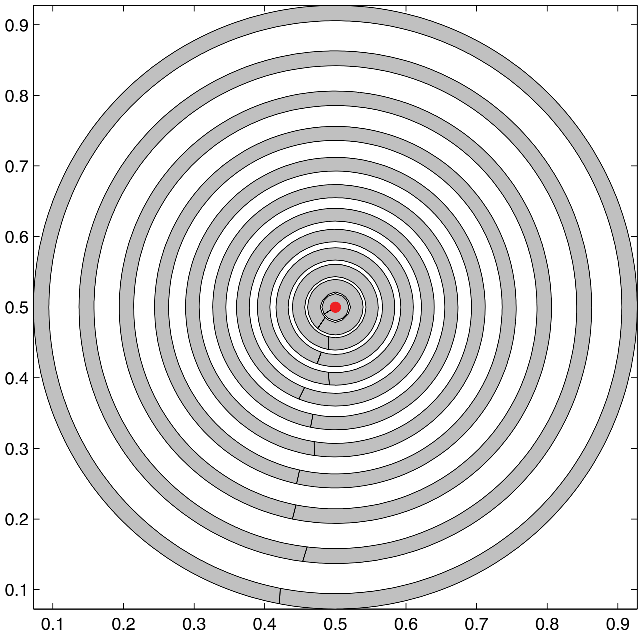


Fig. 6.7. *The calculated polygonal seed layer pattern for the shape given by $F(x,y) = -(x-0.5)^2 - (y-0.5)^2 + 0.25$, shown in Figure 6.6. The dot in the center is the powered point.*

Three values of the number of grid points ($N_x = N_y = 50, 100, 200$) and three values for the number of start times ($n = 10, 20, 40$) are used for the inverse algorithm. More grid points make the boundaries of the layers smoother, and more start

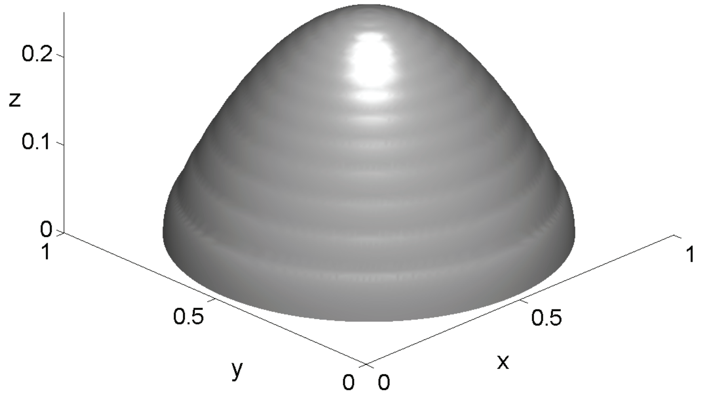


Fig. 6.8. *The forward growth of the calculated polygonal seed layer pattern in Figure* 6.7. *It compares well to the desired shape in Figure* 6.6.

Table 6.1

The L^2 errors for different values of $N_x = N_y$ and n for the function $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$.

n	$\|F - F^\phi\|, N_x = 50$	$\|F - F^\phi\|, N_x = 100$	$\|F - F^\phi\|, N_x = 200$
10	0.00425832	0.00423406	0.00422753
20	0.00201249	0.00198800	0.00198200
40	0.00076916	0.00074762	0.00074282

Table 6.2

The L^2 errors for different values of $N_x = N_y$ and n for the function $F(x, y) = -(x - 0.5)^2 - (y - 0.5)^2 + 0.25$.

n	$\|F - F^\phi\|, N_x = 50$	$\|F - F^\phi\|, N_x = 100$	$\|F - F^\phi\|, N_x = 200$
10	0.00234398	0.00231595	0.00230947
20	0.00092564	0.00089596	0.00088921
40	0.00036701	0.00034218	0.00034155

Table 6.3

Runtimes for different values of $N_x = N_y$ and n for the function $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$.

n	Runtime, $N_x = 50$	Runtime, $N_x = 100$	Runtime, $N_x = 200$
10	0.421	2.714	20.159
20	0.762	5.188	38.746
40	1.462	10.195	76.701

times produce more unpowered layers in the seed layer. Tables 6.1 and 6.2 show the computed errors for the functions $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$ and $F(x, y) = -(x-0.5)^2 - (y-0.5)^2 + 0.25$, respectively. As the data indicate, the error has very little dependence on $N_x = N_y$ but has nearly linear dependence on n^{-1}. Table 6.3 shows the run times for the function $F(x, y) = 0.1\cos(\pi x) + 0.25\sin(\pi y) + 0.1$ with different numbers of grid points and different values of n. All computations were performed on a PC with an Intel M1.6GHz processor with 1.25GB of RAM.

Examining Tables 6.1 and 6.3 shows there is no benefit to using more grid points than needed for resolution of the layers. Using a larger n, however, does improve the accuracy of the seed layer, with only a linear increase in computational time.

7. Conclusions and future work.

The method presented here may face material limitations. An accurate fit to a desired shape can result in seed layer elements that are thin strips. The limit on the minimum geometry of the seed layer elements is governed by the photolithographic technology used to pattern them. Typically this ranges from about 1 μm, with readily available optical lithographic systems, to substantially less than 100 nm with electron-beam lithography and industry-leading optical photolithography systems.

Surface roughness could also limit the seed layer geometry. The process used to define the seed layer typically consists of a combination of photolithography, physical vapor deposition (PVD) (e.g., evaporation or sputtering), and chemical etching. The metal films deposited by PVD can be very thin ($<$ 100 nm) and are very smooth. The roughness of electrodeposited films can be highly variable and are subject to the specific process and recipe used. To obtain smoother electrodeposition, one can add chemicals (i.e., brighteners) and perform periodic current reversal.

The solution of the forward problem using the level set method depends on construction of a global distance function from the powered and (initially) unpowered segments. There are several ways to construct this distance function. Our method uses a time discretization that directly mimics the physical evolution of the front. It has some advantages in that additional physics could be easily included. An alternative method in both two and three dimensions would directly construct the distance function from the geometry of the segments.

One additional feature that can be introduced into seed layer patterns is solid boundaries. These solid boundaries are made of an insulated material and prevent growth of the metal shapes beyond the boundaries. This allows even more shape possibilities. Future work would include these solid boundaries in the model for forward growth and would attempt to incorporate them into the inverse procedure. Also, the current model is quite simple, and future work would add more physics into the model, such as nonuniform growth or perhaps some diffusion of the growing surface.

Appendix. The distance function for a line segment.

Let the line segment be given as $[(x_1, y_1), (x_2, y_2)]$. Without loss of generality, we shall assume that $x_1 < x_2$. There are 4 cases, depending on the slope of the line segment, which we denote as $m = \frac{y_2 - y_1}{x_2 - x_1}$.

1. $m = \pm\infty$.

$$d(x,y) = \begin{cases} \sqrt{(x-x_1)^2 + (y-y_b)^2} & \text{if } y < y_b, \\ |x - x_1| & \text{if } y_b \le y \le y_t, \\ \sqrt{(x-x_1)^2 + (y-y_t)^2} & \text{if } y > y_t, \end{cases}$$

where $y_b = \min(y_1, y_2)$ and $y_t = \max(y_1, y_2)$.

2. $m = 0$.

$$d(x,y) = \begin{cases} \sqrt{(x-x_1)^2 + (y-y_1)^2} & \text{if } x < x_1, \\ |y - y_1| & \text{if } x_1 \le x \le x_2, \\ \sqrt{(x-x_2)^2 + (y-y_1)^2} & \text{if } x > x_2. \end{cases}$$

3. $0 < m < \infty$.

$$d(x,y) = \begin{cases} \sqrt{(x-x_2)^2+(y-y_2)^2} & \text{if } y \geq (-\frac{1}{m}(x-x_2)+y_2), \\ \sqrt{(x-x_1)^2+(y-y_1)^2} & \text{if } y \leq (-\frac{1}{m}(x-x_1)+y_1), \\ \sqrt{(x-x_m)^2+(y-y_m)^2} & \text{otherwise,} \end{cases}$$

where $x_m = \left(\frac{1}{m+\frac{1}{m}}\right) * \left(mx_1 + \frac{x}{m} + y - y_1\right)$ and $y_m = m * (x_m - x_1) + y_1$.

4. $-\infty < m < 0$.

$$d(x,y) = \begin{cases} \sqrt{(x-x_2)^2+(y-y_2)^2} & \text{if } y \leq (-\frac{1}{m}(x-x_2)+y_2), \\ \sqrt{(x-x_1)^2+(y-y_1)^2} & \text{if } y \geq (-\frac{1}{m}(x-x_1)+y_1), \\ \sqrt{(x-x_m)^2+(y-y_m)^2} & \text{otherwise,} \end{cases}$$

where $x_m = \left(\frac{1}{m+\frac{1}{m}}\right) * \left(mx_1 + \frac{x}{m} + y - y_1\right)$ and $y_m = m * (x_m - x_1) + y_1$.

REFERENCES

[1] D. Adalsteinsson and J. A. Sethian, *A level set approach to a unified model for etching, deposition, and lithography.* 1. *Algorithms and two-dimensional simulations*, J. Comput. Phys., 120 (1995), pp. 128–144.

[2] D. Adalsteinsson and J. A. Sethian, *A level set approach to a unified model for etching, deposition, and lithography.* 2. 3*-dimensional simulations*, J. Comput. Phys., 122 (1995), pp. 348–366.

[3] D. Adalsteinsson and J. A. Sethian, *A level set approach to a unified model for etching, deposition, and lithography.* 3. *Redeposition, reemission, surface diffusion, and complex simulations*, J. Comput. Phys., 138 (1997), pp. 193–223.

[4] S. Chen, M. Kang, B. Merriman, R. E. Caflisch, C. Ratsch, R. Fedkiw, M. F. Gyure, and S. Osher, *Level set method for thin film epitaxial growth*, J. Comput. Phys., 167 (2001), pp. 475–500.

[5] D. Josell, D. Wheeler, W. H. Huber, J. E. Bonevich, and T. P. Moffat, *A simple equation for predicting superconformal electrodeposition in submicrometer trenches*, J. Electrochem. Soc., 148 (2001), pp. C767–C773.

[6] P. Limousin, P. Krack, P. Pollak, A. Benazzouz, C. Ardouin, D. Hoffmann, and A.-L. Benabid, *Electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease*, New England J. Medicine, 339 (1998), pp. 1105–1111.

[7] P. S. Motta and J. W. Judy, *Multielectrode microprobes for deep brain stimulation fabricated with a customizable* 3*-D electroplating process*, IEEE Trans. Biomed. Engrg., 52 (2005), pp. 923–933.

[8] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York, 2003.

[9] D. Wheeler, D. Josell, and T. P. Moffat, *Modeling superconformal electrodeposition using the level set method*, J. Electrochem. Soc., 150 (2003), pp. C302–C310.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 1065–1083

ACOUSTIC WAVES IN LONG RANGE RANDOM MEDIA[*]

RENAUD MARTY[†] AND KNUT SOLNA[‡]

Abstract. We consider waves propagating through multiscale media. Much is known about waves propagating through a medium that satisfies a scale separation assumption with random fluctuations on a microscale. Here we go beyond this situation and consider waves propagating through a medium defined in terms of a long range process. Such a medium can, for instance, be modeled in terms of a one-dimensional fractional Brownian motion with variations on a continuum of scales. Fractal medium models are used to model, for example, the heterogeneous earth and the turbulent atmosphere. We set forth a framework using the theory of rough paths in which propagation problems of this nature can be analyzed in the case with anticipative medium fluctuations with a Hurst exponent $H > 1/2$. We show how the wave interacts with the medium fluctuations in this case and that the interaction is qualitatively different from the situation where the medium satisfies a separation of scales assumption. In the long range case considered here the travel time depends strongly on the particular medium realization, but in fact the pulse shape does not.

Key words. wave propagation, random media, long range processes, fractional Brownian motion

AMS subject classifications. 34F05, 34E10, 37H10, 60H20

DOI. 10.1137/07068610X

1. Introduction.

Modeling in terms of a multiscale medium is important for propagation problems in, for instance, the earth's crust, the turbulent atmosphere, turbulent boundary layers, sea ice, and outer space [9, 15, 19, 20, 34, 42, 44]. Communication, remote sensing, and laser beam propagation schemes are affected by bad weather and multiscale medium variations. Large scale research projects (the ABLE ACE program Kirtland AFB, for instance [43]) have focused on gathering atmospheric turbulence data and numerically simulating propagation of wave fields through synthetic turbulence models that derive from these. Above a boundary layer atmospheric turbulence may occur within a stratified environment, and the turbulent temperature variations may be highly anisotropic; see [16, 36]. Rough and long range medium fluctuations associated with multiscale modeling are also important in medical imaging, device modeling, and nuclear technology, to name a few. The zone in between different tissue types (or in between different dielectrica) may in particular be strongly heterogeneous with variation on a continuum of scales. In general, the detailed pointwise variation of a multiscale medium, the refractive index, say, cannot be identified. However, the statistics of this variation can be characterized. Optimal design of, for instance, imaging and communication algorithms requires insight about how the wave is affected by the rough medium fluctuations, that is, the nonlinear coupling between medium and wave field statistics. This is particularly the case with modern high resolution sampling and imaging technology. Insight about the wave medium interaction is also important in a range of other applications like design of sound-absorbing materials and nondestructive evaluation of fractured materials. A good understanding

[*] Received by the editors March 22, 2007; accepted for publication (in revised form) September 12, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/siap/69-4/68610.html

[†] Institut Elie Cartan Nancy, Nancy-Université, CNRS, INRIA, Boulevard des Aiguillettes, B.P. 239, F-54506 Vandœuvre lès Nancy, France (renaud.marty@iecn.u-nancy.fr).

[‡] Department of Mathematics, University of California at Irvine, Irvine, CA 92697 (ksolna@math.uci.edu). Work partially supported by DARPA grant N00014-02-1-0603, ONR grant N00014-02-1-0090, NSF grant 0307011, and the Sloan Foundation.

of how the wave interacts with variations on a continuum of small length scales is therefore important.

Propagation of high frequency waves in *smooth* media is well understood. A lot is also known about propagation in heterogeneous media that vary on a well-defined microscale. However, propagation in *rough* or *multiscale* media is not so well understood. We will look at how propagating pulses interact with rough variations in the medium. In [10, 11, 12] we considered propagation in rough media when the wave phenomenon was modeled in terms of the paraxial or forward approximation. In [37] we considered the full wave equation and a discrete multiscale medium. Here, we continue this line of research by analyzing the full wave equation in the context of a one-dimensional continuous multiscale medium modeled in terms of a long range process with slowly decaying correlations, and we consider in particular fractional Brownian motion–based media.

In the homogenization or effective medium regime, with the width of the propagating pulse being large compared to the scale of the medium fluctuations and propagation distances on the scale of the wavelength, the rapidly varying properties of the medium can be replaced by their homogenized or averaged values. However, over long propagation distances the accumulated effect of the scattering, associated with the medium microstructure, gradually changes the pulse *beyond* the geometrical effects of the high frequency analysis in the smooth homogenized medium. These modifications depend in general on the particular medium realization. Thus, to describe the propagation phenomenon it is not enough to consider only the mean wave field; one should also aim at describing the character of the fluctuations in the wave field. A mathematical theory for pulse propagation has been developed in [1, 6, 14, 25]. It deals with pulses in a particular realization of the random medium, and it explains why in many cases the evolution of the *pulse shape* is to leading order *deterministic*. We refer to this phenomenon as pulse *stabilization*. So far, two salient features of this "pulse shaping" theory have been that it assumes a one-dimensional medium and a separation of scales for the medium heterogeneities, that is, that the medium has features on microscales which are well separated from the macroscale. However, as explained above, many empirical studies suggest that, for instance, the earth's crust should be modeled as containing fluctuations on a continuum of length scales. Stabilization and pulse shaping in a two scale medium with slow lateral variations in the medium has been analyzed in [38]. Here, we generalize the pulse shaping theory for a two scale medium to the long range multiscale case.

We analyze acoustic waves propagating in a one-dimensional medium, modeled in terms of a long range process. As a particular example we consider media defined in terms of fractional Brownian motion. Fractional Brownian motion is a Gaussian (self-similar) stochastic process and is often used as a model for processes containing fluctuations on a continuum of length scales, for instance for modeling of turbulent environments. The Hurst exponent H characterizes the roughness of the fractional Brownian motion, and the value $H = 1/2$ gives standard Brownian motion. In the simplest case with $H = 1/2$ the medium model that we consider satisfies a separation of scales assumption. For $H \neq 1/2$ the medium contains long range interactions and variations on many scales. In this case the correlations in the medium have only polynomial decay, and our objective is to analyze the effects such long range correlations have on the propagating wave. Here, we shall analyze the case with $H > 1/2$ corresponding to persistent fluctuations so that consecutive increments of the process are positively correlated [13]. In fact, we shall show that the pulse shape is *not* affected by the random medium fluctuations to leading order. However, the travel time of the pulse depends on the particular medium realization, and the travel time fluctuations are large relative to the period of the pulse.

A number of studies of wave interaction with a fractal or multiscale object deal with scattering caused by fractal interfaces. However, some authors have explored wave-interaction with deterministic fractal media using numerical simulations [5, 21, 39]. Here, we present a mathematical analysis of acoustic pulse transmission through a random fractal and illustrate our theoretical results with numerical simulations. In section 2 we introduce the problem and review the basic wave decomposition approach and the classical scale separation result. Next, in section 3 we introduce the multiscale medium and summarize how the pulse shaping theory generalizes to these media. In section 4, we illustrate our theoretical results with numerical simulations. Finally, section 5 is devoted to the derivation of the main result.

2. Wave decomposition. The governing equations are the Euler equations giving conservation of moments and mass:

$$\rho(z)\frac{\partial u}{\partial t}(z,t)+\frac{\partial p}{\partial z}(z,t)=0\,, \tag{2.1}$$

$$\frac{1}{K(z)}\frac{\partial p}{\partial t}(z,t)+\frac{\partial u}{\partial z}(z,t)=0\,, \tag{2.2}$$

where t is the time, z is the depth into the medium, p is the pressure, and u the particle velocity. The medium parameters are the density ρ and the bulk-modulus K (reciprocal of the compressibility). We assume that ρ is a constant identically equal to one in our nondimensionalized units and that $1/K$ is modeled as follows:

$$\frac{1}{K(z)}=\begin{cases} 1+\varepsilon^{\kappa}\nu\left(\dfrac{z}{\varepsilon^2}\right) & \text{for } z\in[0,Z]\,,\\ 1 & \text{for } z\in\mathbb{R}-[0,Z]\,, \end{cases} \tag{2.3}$$

where $\kappa\geq 0$. We introduce the right- and left-going waves

$$A=p+u \quad \text{and} \quad B=u-p\,, \tag{2.4}$$

where the boundary conditions are of the form

$$A(z=0,t)=f(t/\varepsilon^{\tau}) \quad \text{and} \quad B(z=Z,t)=0\,, \tag{2.5}$$

for a positive real number $\tau>0$ and a source function f. In order to deduce a description of the transmitted pulse, we open a window of size ε^{τ} in the neighborhood of the travel time of the homogenized medium and define the processes

$$a^{\varepsilon}(z,s)=A(z,z+\varepsilon^{\tau}s) \quad \text{and} \quad b^{\varepsilon}(z,s)=B(z,-z+\varepsilon^{\tau}s)\,. \tag{2.6}$$

Observe that the background or homogenized medium in our scaling has a constant speed of sound equal to unity and that the medium is matched so that in the frame introduced in (2.6) the pulse-shapes of the right- and left-going waves are constant in the slab if $\nu\equiv 0$ or if we consider the homogenized medium [14]. We introduce next the Fourier transforms $\widehat{a}^{\varepsilon}$ and $\widehat{b}^{\varepsilon}$ of a^{ε} and b^{ε}, respectively,

$$\widehat{a}^{\varepsilon}(z,\omega)=\int e^{i\omega s}a^{\varepsilon}(z,s)ds \quad \text{and} \quad \widehat{b}^{\varepsilon}(z,\omega)=\int e^{i\omega s}b^{\varepsilon}(z,s)ds\,,$$

that satisfy

$$\frac{d\widehat{a}^\varepsilon}{dz} = \frac{i\omega}{2\varepsilon^{\tau-\kappa}}\nu\left(\frac{z}{\varepsilon^2}\right)\left(\widehat{a}^\varepsilon - e^{-2i\omega z/\varepsilon^\tau}\widehat{b}^\varepsilon\right), \qquad \widehat{a}^\varepsilon(0,\omega) = \widehat{f}(\omega)\,, \tag{2.7}$$

$$\frac{d\widehat{b}^\varepsilon}{dz} = \frac{i\omega}{2\varepsilon^{\tau-\kappa}}\nu\left(\frac{z}{\varepsilon^2}\right)\left(e^{2i\omega z/\varepsilon^\tau}\widehat{a}^\varepsilon - \widehat{b}^\varepsilon\right), \qquad \widehat{b}^\varepsilon(Z,\omega) = 0\,. \tag{2.8}$$

Following [6, 14], we express the previous system of equations in term of propagator $P^\varepsilon_\omega(z)$, which can be written as

$$P^\varepsilon_\omega(z) = \begin{pmatrix} \alpha^\varepsilon_\omega(z) & \overline{\beta^\varepsilon_\omega}(z) \\ \beta^\varepsilon_\omega(z) & \overline{\alpha^\varepsilon_\omega}(z) \end{pmatrix}, \tag{2.9}$$

and which satisfies

$$\frac{dP^\varepsilon_\omega}{dz}(z) = \frac{1}{\varepsilon^\gamma}H_\omega\left(\frac{z}{\varepsilon^\tau}, \frac{z}{\varepsilon^2}\right)P^\varepsilon_\omega(z)\,, \qquad P^\varepsilon_\omega(z=0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.10}$$

with $\gamma = \tau - \kappa$ and

$$H_\omega(z_1, z_2) = \frac{i\omega}{2}\nu(z_2)\begin{pmatrix} 1 & -e^{-2i\omega z_1} \\ e^{2i\omega z_1} & -1 \end{pmatrix}.$$

Defining next the transmission coefficient T^ε_ω and the reflection coefficient R^ε_ω by

$$T^\varepsilon_\omega(z) = \frac{1}{\overline{\alpha^\varepsilon_\omega}(z)} \quad \text{and} \quad R^\varepsilon_\omega(z) = \frac{\beta^\varepsilon_\omega(z)}{\overline{\alpha^\varepsilon_\omega}(z)}\,, \tag{2.11}$$

we can write

$$a^\varepsilon(Z,s) = \frac{1}{2\pi}\int e^{-is\omega}T^\varepsilon_\omega(Z)\widehat{f}(\omega)\,d\omega \tag{2.12}$$

and

$$b^\varepsilon(0,s) = \frac{1}{2\pi}\int e^{-is\omega}R^\varepsilon_\omega(Z)\widehat{f}(\omega)\,d\omega\,. \tag{2.13}$$

Henceforth, we shall study the asymptotics of the propagator P^ε_ω in order to characterize a^ε and b^ε as ε goes to 0.

We recall now what happens in a "short range" model when $\tau = 1$ and $\kappa = 0$. We assume that ν is a centered Markov process with an invariant probability measure whose generator satisfies the Fredholm alternative. It is well known [6, 14] that under these assumptions, in order to characterize the transmitted pulse, the propagator equations P^ε_ω can be replaced by an effective system of stochastic differential equations from which we can deduce that, as ε goes to 0,

$$a^\varepsilon(Z,s) \longrightarrow \widetilde{a}(Z,s)\,, \tag{2.14}$$

with

$$\widetilde{a}(Z,s) = (f * G)(s - B)\,, \tag{2.15}$$

where G is a centered Gaussian density and B a Gaussian random variable. Proving this result involves using the diffusion approximation theorem [14] to get asymptotic expressions for specific propagator moments from which we can deduce the expression for the limit $\widetilde{a}(Z, s)$. Therefore, when we capture the pulse at its random arrival time we will see a pulse whose shape does not depend on the realization of the random medium. This shape is the original pulse shape convolved with a Gaussian pulse shaping kernel. Thus, the effects of the random medium fluctuations can be described in terms of a random travel time correction and an anomalous diffusion effect. By the same approach we can also prove that

$$b^{\varepsilon}(0, s) \longrightarrow 0\,. \tag{2.16}$$

This result shows that in the case with a constant homogenized medium the reflected wave will be negligible in the small ε limit. While the mathematical derivation of this "medium pulse shaping" result was first obtained in [6] and [26], it was first derived in the geophysical literature in [32] via a heuristic derivation. The approximation has since come to be known as the O'Doherty–Anstey (ODA) theory after the authors of the pioneering paper, which lead to a string of papers both in the mathematical and geophysical literature, reflecting its relevance. However, so far the problem has only been analyzed in short range media. Here we generalize to the situation with long range media and show that then we get a qualitatively different result. We explain in section 3.2 that in this case we have a strong travel time perturbation, as in the classic theory; however, in the case with long range media the pulse shape itself is *stable.*

3. Medium model and main result.

In this section we investigate the propagation in a long range medium. We first describe the model in detail, and then we establish the main result of the paper.

3.1. Long range model.

We assume $\gamma \in (0, 1)$ and that ν can be written as $\nu(z) = T(m(z))$ for every z where the following hold:

- T is a continuous function which is strictly bounded by 1 in absolute value, odd, and increasing. Note that our analysis remains valid in the case with T not being bounded; the boundedness by 1 is introduced to make the model physically pertinent: we recall that $1 + \varepsilon^{\kappa}\nu$ is the compressibility with $\kappa \geq 0$.
- m is a Gaussian process, centered, stationary and has a correlation function r_m which has the following asymptotic property as z goes to ∞:

$$r_m(z) = \mathbb{E}[m(0)m(z)] \sim c_m z^{-\gamma}\,. \tag{3.1}$$

Note that therefore the medium fluctuations ν themselves are not Gaussian in general; their distribution is controlled via the choice of T.

The property (3.1) is the so-called long range property. Its main consequence is that the covariance function r_m of m is not absolutely integrable:

$$\int_0^{\infty} |r_m(z)|\, dz = \infty\,.$$

Hence, this situation is in dramatic contrast with the classical mixing (or short range) case. Indeed, a mixing process has an integrable covariance function [22]. Another important consequence of (3.1) regards the choice of scales. We are given the correlation length of the medium $(1/\varepsilon^2)$, the amplitude (ε^{κ}), and the rate of decorrelation

($r_m(z) \sim c_m/z^\gamma$) of the random perturbations. Then, because our goal is to capture the behavior of the transmitted wave pulse and its interaction with the medium, we have to choose an appropriate size of the window to capture the critical wavelength interaction scale by taking $\tau = \gamma + \kappa$. We next explain this scaling choice. The propagator equation is essentially driven by the process $\varepsilon^{\kappa-\tau} m\left(z/\varepsilon^2\right)$, and, we shall see that indeed its analysis involves the study of the convergence of the antiderivative w^ε of this process:

$$w^\varepsilon(z) = \int_0^z \frac{1}{\varepsilon^{\tau-\kappa}} m\left(\frac{z'}{\varepsilon^2}\right) dz' .$$

However, it is known [35], as recalled in Lemma 1 just below, that the appropriate scale then is $\gamma = \tau - \kappa$. This gives convergence to a fractional Brownian motion. Let $H \in (0,1)$; then fractional (one-dimensional) Brownian motion (fBm) with Hurst parameter H is the centered Gaussian process $(B_H(z))_{z\in\mathbb{R}}$ with covariance function

$$\mathbb{E}[B_H(z_1)B_H(z_2)] = \frac{1}{2}\{|z_1|^{2H} + |z_2|^{2H} - |z_1 - z_2|^{2H}\} .$$

We refer the reader to Samorodnitsky and Taqqu's book [35] for a good reference. For the process w^ε we have the convergence given next.

Lemma 1. *Let $H = (2-\gamma)/2$. As ε goes to 0, the finite-dimensional distributions of w^ε converge to those of the fractional Brownian motion $c'_H B_H$, where ${c'_H}^2 = c_m H^{-1}(2H-1)^{-1}$.*

Now we give two examples of processes m:

- Fractional white noise with Hurst parameter $H = (2-\gamma)/2 \in (1/2, 1)$ that can be defined as

  $$m(z) = B_H(z+1) - B_H(z) , \tag{3.2}$$

  where B_H is the fBm with Hurst parameter H.
- The (stationary) fractional Ornstein–Uhlenbeck process with index $H = (2-\gamma)/2$ defined by

  $$m(z) = B_H(z) - e^{-z}\int_{-\infty}^z e^{z'} B_H(z')\, dz' , \tag{3.3}$$

  where B_H is the fBm with Hurst parameter H. As in the case with fractional white noise with index H, the fractional Ornstein–Uhlenbeck process is continuous, Gaussian, stationary, and centered and satisfies (3.1).

Notice that here we presented two examples for m in terms of an fBm, but all results in this paper are true and proved under the general assumptions on m presented above, in particular Gaussianity and slow correlation decay. Notice also that we shall consider here

$$H \in (1/2, 1) \quad \text{so that} \quad \gamma \in (0,1) . \tag{3.4}$$

We next introduce some notation. We denote by X a Gaussian, centered, and reduced random variable: $X \sim \mathcal{N}(0,1)$. Letting $\sigma_0 = \sqrt{\mathbb{E}[m(0)^2]}$, we will need the Hermite development of the function $T(\sigma_0 \times \cdot)$. We denote by H_k the kth Hermite polynomial and by $J(k)$ the kth Hermite coefficient of the function $T(\sigma_0 \times \cdot)$, that is to say, $J(k) = \mathbb{E}[T(\sigma_0 X)H_k(X)]$. Thanks to the assumptions on T, we have $J(0) = 0$

and $J(1) \neq 0$, so that the Hermite coefficient of T is 1. Therefore, we can write (see the appendix for more details about Hermite polynomials)

$$T(\sigma_0 \times X) = \sum_{k=1}^{\infty} \frac{J(k)}{k!} H_k(X) \,. \tag{3.5}$$

We conclude this subsection by establishing the long range behavior of ν.

LEMMA 2. *For $z \to \infty$ we have*

$$r_\nu(z) := \mathbb{E}[\nu(0)\nu(z)] \sim \frac{c_\nu}{z^\gamma} \,,$$

where $c_\nu = c_m J(1)^2/\sigma_0^2 = c_m \mathbb{E}[XT(\sigma_0 X)]^2/\sigma_0^2$.

Proof. In view of (3.5) we can write (using (A.1))

$$\nu(z) = \sum_{k=1}^{\infty} \frac{J(k)}{k!} H_k\left(\frac{m(z)}{\sigma_0}\right) \,, \tag{3.6}$$

so that (using (A.3))

$$\begin{aligned}\mathbb{E}[\nu(0)\nu(z)] &= \sum_{k=1}^{\infty} \frac{J(k)^2}{(k!)^2} \mathbb{E}\left[H_k\left(\frac{m(0)}{\sigma_0}\right) H_k\left(\frac{m(z)}{\sigma_0}\right)\right] \\ &= \sum_{k=1}^{\infty} \frac{J(k)^2}{k!\sigma_0^{2k}} r_m(z)^k \,.\end{aligned}$$

Therefore, we need to study the limit of

$$z^\gamma \mathbb{E}[\nu(0)\nu(z)] = \sum_{k=1}^{\infty} \frac{J(k)^2}{k!\sigma_0^{2k}} z^\gamma r_m(z)^k \,.$$

Observe that for $k = 1$ we have $z^\gamma r_m(z) \sim c_m$ as $z \to \infty$, and for $k > 1$ we have $z^\gamma r_m(z)^k \to 0$. Moreover, we have the uniform upper bound for z sufficiently large:

$$\frac{J(k)^2}{k!\sigma_0^{2k}} z^\gamma |r_m(z)|^k \leq \frac{J(k)^2}{k!} \,.$$

Using the fact that (by (A.2))

$$\sum_{k=1}^{\infty} \frac{J(k)^2}{k!} < \infty \,, \tag{3.7}$$

the result now follows from the uniform convergence theorem. □

3.2. Main result. Now we establish the main result of this paper. We shall see that the asymptotic behavior of the transmission coefficient is quite different in the long range case than in the short range case. Recall that we let $\tau = \gamma + \kappa > 0$.

THEOREM 1. *Under the above assumptions, as ε goes to 0, $\{a^\varepsilon(Z,s)\}_s$ converges in distribution in the space of continuous functions endowed with the uniform topology to the random process $\{\widetilde{a}(Z,s)\}_s$ that can be written as*

$$\widetilde{a}(Z,s) = f\left(s - \frac{c_H}{2} B_H(Z)\right) \,, \tag{3.8}$$

where B_H is a fractional Brownian motion with Hurst parameter $H = (2 - \gamma)/2$ and $c_H^2 = c_\nu H^{-1}(2H - 1)^{-1}$ with c_ν as introduced in Lemma 2. Moreover, the process $\{b^\varepsilon(0, s)\}_s$ converges to 0.

Therefore, we see that the pulse is stable and does not undergo a deterministic evolution in time as in the short range case. In the short range case the evolution of the pulse shape and the randomization of the pulse travel-distance takes place on the same time scale. In the long range case with persistent medium fluctuations and slow decorrelation the evolution of the pulse shape takes place on a relatively slow time scale, and in Theorem 1 we observe only the randomization of the travel distance while the pulse shape is stable. In the long range case with $H > 1/2$ the medium fluctuations are persistent and "smoother" than in the classic case, so that the traveltime perturbation corresponding to an accumulation of fluctuations effect becomes relatively stronger than the pulse transformation effect which is due to scattering and enhanced by the roughness of the medium. Note that the traveltime perturbation is in the long range case described by a fractional Brownian motion with a Hurst index that corresponds to the effective Hurst index for the medium perturbations, while in the classic case it is described by a standard Brownian motion.

From the point of view of modern applications of the theory of waves in random media the above result is relevant. Recently there has been a lot of interest in imaging schemes in the context of cluttered layered media [3, 14] exploiting the ODA approximation, for instance, as well as in nonlayered media [4]. This reflects the fact that classic imaging schemes deteriorates when the "background" medium becomes fluctuating. The above results show how this body of results applies to the long range situation, in which case the modeling of the travel time perturbation becomes the important aspect. We remark also that currently there is a lot of interest in the design of robust wireless communication schemes when the signalling takes place through clutter, through the turbulent atmosphere, for instance, which is relevant also in the context of remote sensing. Design of robust schemes requires a forward model that captures the interaction of the pulse with the medium, which in the long range persistent and layered case is described by Theorem 1. The derivation of this result, which is presented in section 5, sets forth a framework which we expect will be useful also in a more general context to describe other physical scaling scenarios, analogous to the case with short range medium fluctuations [17, 18].

4. Numerical illustration. We illustrate the results with some numerical simulations. In the numerical simulations we use a Gaussian initial pulse shape. In the normalized coordinates the support of the initial pulse is 10^{-3} and the total propagation distance is 1. The medium is defined as in (3.2) with $\kappa = 2H$ and, moreover, with $\epsilon = 10^{-2}$ and with a cutoff function that is the identity in the neighborhood of the origin with a smooth cutoff. We use a discretization corresponding to equal travel time sections and the method described in [33] to simulate the realizations of the medium fluctuations. In Figure 4.1 we show the result of three simulations when the propagated pulse is plotted relative to its random arrival time when $H = .6$ and on the fine scale ε^τ. Observe that indeed the pulse shape is to leading order not affected by random fluctuations in the medium, as predicted by Theorem 1. In Figure 4.2 we show the corresponding picture when $H = 1/2$. Note that in this case the pulse shape is modified via a convolution with a Gaussian kernel as described by the classical pulse shaping or ODA theory in the case of strong mixing.

5. Proof of Theorem 1. We first give an outline of the proof. As recalled in section 2 the process $\{a^\varepsilon(Z, s)\}_s$ can be written in terms of the propagator P^ε_ω, and

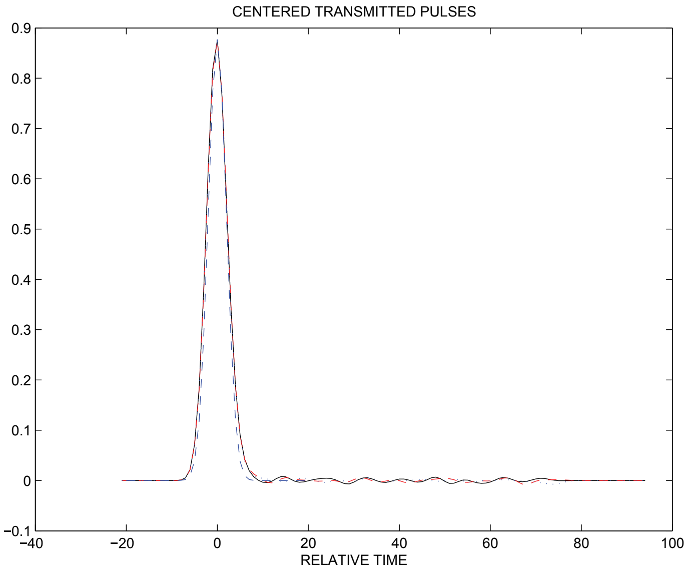


FIG. 4.1. *The transmitted wave shown at a fixed depth for several medium realizations and with $H = .6$. The dashed line is the original pulse shape.*

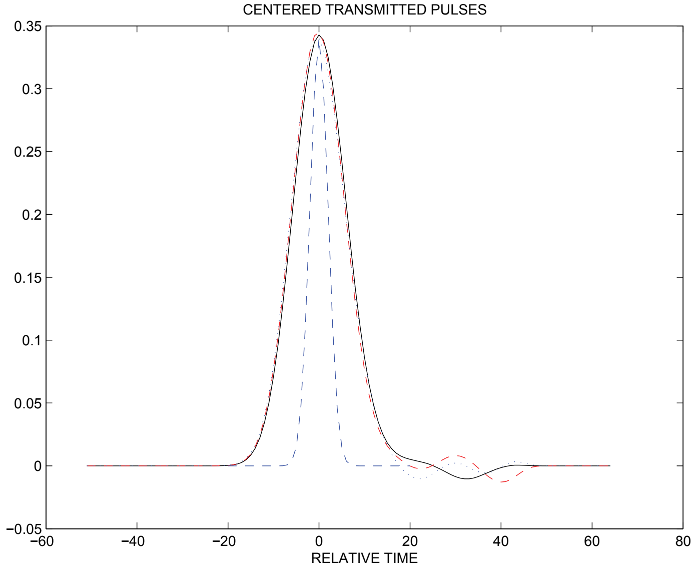


FIG. 4.2. *The transmitted wave shown at a fixed depth for several medium realizations and with $H = 1/2$. The dashed line is the original pulse shape.*

thus the study of the convergence of $\{a^\varepsilon(Z, s)\}_s$ can be analyzed via asymptotic properties of P^ε_ω. This convergence analysis will follow the lines of [30, 31]. The propagator P^ε_ω satisfies the equation

$$\frac{dP^\varepsilon_\omega}{dz}(z) = \frac{1}{\varepsilon^\gamma} H_\omega\left(\frac{z}{\varepsilon^\tau}, \frac{z}{\varepsilon^2}\right) P^\varepsilon_\omega(z)\,,$$

which we can write in the form

$$dP^\varepsilon_\omega(z) = \frac{i\omega}{2}\sum_{j=1}^{3} F_j P^\varepsilon_\omega \, dv^\varepsilon_j(z)\,, \tag{5.1}$$

where

$$F_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad F_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and v^ε_1, v^ε_2, and v^ε_3 are three processes of bounded variation that we can write as

$$\begin{aligned} v^\varepsilon_1(z) &= \int_0^z \frac{1}{\varepsilon^\gamma}\nu\left(\frac{z'}{\varepsilon^2}\right) dz'\,, \\ v^\varepsilon_2(z) &= \int_0^z \frac{1}{\varepsilon^\gamma}\nu\left(\frac{z'}{\varepsilon^2}\right)\cos\left(2\omega\frac{z'}{\varepsilon^\tau}\right) dz'\,, \\ v^\varepsilon_3(z) &= \int_0^z \frac{1}{\varepsilon^\gamma}\nu\left(\frac{z'}{\varepsilon^2}\right)\sin\left(2\omega\frac{z'}{\varepsilon^\tau}\right) dz'\,. \end{aligned}$$

Thanks to Lyons's rough paths theory for which we recall some tools in section 5.1, we shall see that the convergence of P^ε_ω can be reduced for a convenient topology to the convergence of the process $\mathbf{v}^\varepsilon$ defined as

$$\mathbf{v}^\varepsilon := (v^\varepsilon_1, v^\varepsilon_2, v^\varepsilon_3)\,.$$

Hence, we first prove the convergence of $\mathbf{v}^\varepsilon$, then by Theorem 2 below we deduce the convergence of P^ε_ω, and thanks to (2.12) we finally conclude by the convergence of $\{a^\varepsilon(Z,s)\}_s$.

5.1. Rough paths.

In this section we fix $p \in [1,2)$ and consider a closed interval $I = [0, Z]$. We define the p-variation of a continuous function $w : I \to \mathbb{R}^n$ by

$$V_p(w) := \left(\sup_D \sum_{j=0}^{k-1} \|w(z_{j+1}) - w(z_j)\|^p\right)^{1/p},$$

where $\sup_D$ runs over all finite partitions $\{0 = z_0, \ldots, z_k = Z\}$ of I and where here and below $\|\cdot\|$ refers to the Euclidean norm. The space of all continuous functions of bounded variation (1-variation) is endowed with the p-variation distance

$$\|w\|_p = V_p(w) + \sup_{z\in[0,1]} |w(z)|$$

and is denoted by Ω^∞_p. The closure of this metric space is called the space of all geometric rough paths and is denoted by Ω_p. One of the most important theorems of rough paths theory is the following.

THEOREM 2 (Lyons's continuity theorem [27]). *Let*[1] $G : \mathbb{R}\times\mathbb{R}^d \to \mathcal{L}(\mathbb{R},\mathbb{R}^d)$ *and* $F : \mathbb{R}\times\mathbb{R}^d \to \mathcal{L}(\mathbb{R}^n,\mathbb{R}^d)$ *be two smooth functions. Let y be the unique solution of the differential equation*

$$dy(z) = G(z, y(z))\,dz + F(z, y(z))\,dw(z), \qquad y(z=0) = y_0\,,$$

[1] Here $\mathcal{L}(\mathbb{R},\mathbb{R}^d)$ (resp., $\mathcal{L}(\mathbb{R}^n,\mathbb{R}^d)$) denotes the space of all linear maps from $\mathbb{R}$ (resp., $\mathbb{R}^n$) to $\mathbb{R}^d$.

where w is a bounded variation function. Then the Itô map $\mathcal{I} : w \mapsto y$ is continuous with respect to the p-variation distance from $\Omega_p^\infty(\mathbb{R}^n)$ to $\Omega_p^\infty(\mathbb{R}^d)$. Therefore there exists a unique extension of this map (that we still denote by $\mathcal{I}$) to the space $\Omega_p(\mathbb{R}^n)$.

This theorem has been proved by Lyons and extensively studied and applied (see [8, 23, 24, 27, 28, 29]).

The proof of Theorem 1 is based on analysis of the tightness in the space of geometric rough paths. In the context of this we need to compute the p-variation for $p > 1$. To this effect we will need the following lemmas, of which the first can be found, for instance, in [24] and the second in [23, 24].

Lemma 3. *Let $q \in [1, 2)$ and $(v^\varepsilon)_{\varepsilon>0}$ be a family of continuous random processes of finite q-variation which is tight in the space of continuous functions on I and satisfies*

$$\lim_{A\to+\infty} \sup_{\varepsilon>0} \mathbb{P}[V_q(v^\varepsilon) > A] = 0\,. \tag{5.2}$$

Then $(v^\varepsilon)_{\varepsilon>0}$ is tight in Ω_p for every $p > q$.

Lemma 4. *For every $n \in \mathbb{N}$ and every $k = 0, 1, \ldots, 2^n$, we let $z_k^n := Zk/2^n$. Let $q \in [1, 2)$ and v be a function of finite q-variation. Then there exist two positive constants C_1, C_2 which do not depend on v such that*

$$V_q(v)^q \le C_1 \sum_{n=1}^{+\infty} n^{C_2} \sum_{k=1}^{2^n} \|v(z_k^n) - v(z_{k-1}^n)\|^q\,.$$

We conclude this subsection by mentioning an application of this theory to fractional Brownian motion introduced in section 3.1. From the definition of fBm B_H with index H we remark that if $H = 1/2$, the process B_H is the classical Brownian motion (cBm). However, if $H \ne 1/2$, W_H is neither a semimartingale nor a Markov process. As a consequence, the construction for the fBm of a stochastic calculus turns out to be more involved than for the cBm. This can be done by several way [7], and here we use the rough paths approach as in [8].

5.2. Convergence of the propagator. Using Theorem 2 and the expression (5.1), the asymptotic study of the propagator is reduced to finding the limit in a rough path space of $\mathbf{v}^\varepsilon := (v_1^\varepsilon, v_2^\varepsilon, v_3^\varepsilon)$. This is the subject of the following lemma.

Lemma 5. *Let $p > 2/(2-\gamma) \equiv 1/H$. As ε goes to 0, the increments of $\mathbf{v}^\varepsilon$ converge in Ω_p to those of W_H, which can be written as*

$$W_H = (c_H B_H, 0, 0),$$

where B_H is a fractional Brownian motion with Hurst parameter $H = (2 - \gamma)/2$.

The proof of Lemma 5 is based on establishing several technical lemmas that we derive next. Below we will repeatedly use the notation

$$w^\varepsilon(z) = \int_0^z \frac{1}{\varepsilon^\gamma} m\left(\frac{z'}{\varepsilon^2}\right) dz'\,.$$

We consider first v_1^ε.

Lemma 6. *As ε goes to 0, the finite-dimensional distributions of v_1^ε converge to those of the fractional Brownian motion $c_H B_H$ with ${c_H}^2 = c_\nu H^{-1}(2H - 1)^{-1}$.*

Lemma 6 is a continuous version of [40]. Moreover, a stronger version of this result was established in [41]. Nevertheless, we present here a (simple) proof of Lemma 6 for the sake of completeness.

Proof. In view of Lemma 1 it is enough to show that

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\left|v_1^\varepsilon(z)-\sqrt{\frac{c_\nu}{c_m}}w^\varepsilon(z)\right|^2\right]=0\,. \tag{5.3}$$

Consider the development of $T(\sigma_0\times\cdot)$ in the base of Hermite polynomials $\{H_j\}_{j=0,1,\dots}$,

$$T(m)=\sum_{j=1}^{\infty}\frac{J(j)}{j!}H_j\left(\frac{m}{\sigma_0}\right)\,,$$

and observe that $J(1)/\sigma_0=\sqrt{c_\nu/c_m}$; by (A.3) we then have

$$\begin{aligned}
\mathbb{E}\left[\left|v_1^\varepsilon(z)-\sqrt{\frac{c_\nu}{c_m}}w^\varepsilon(z)\right|^2\right]&=\mathbb{E}\left[\left|\varepsilon^{2H}\int_0^{z/\varepsilon^2}\sum_{j=2}^{\infty}\frac{J(j)}{j!}H_j\left(\frac{m(x)}{\sigma_0}\right)dx\right|^2\right]\\
&=\varepsilon^{4H}\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy\sum_{j=2}^{\infty}\left(\frac{J(j)}{j!}\right)^2\mathbb{E}\left[H_j\left(\frac{m(x)}{\sigma_0}\right)H_j\left(\frac{m(y)}{\sigma_0}\right)\right]\\
&=\varepsilon^{4H}\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy\sum_{j=2}^{\infty}\frac{J(j)^2}{j!}\left(\frac{r(x-y)}{\sigma_0^2}\right)^j\\
&\le\varepsilon^{4H}\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy\sum_{j=2}^{\infty}\frac{J(j)^2}{j!}\left(\frac{r(x-y)}{\sigma_0^2}\right)^2\\
&\le\varepsilon^{4H}C'\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy\,r(x-y)^2\,,
\end{aligned}$$

with

$$C'=\sum_{j=2}^{\infty}\frac{J(j)^2}{j!\sigma_0^4}<\infty\,.$$

As $u\to\infty$, we have $r(u)\sim cu^{-\gamma}$; therefore for every $\eta>0$ there exist z_η, C_η, and $\tilde{C}_\eta$ such that

$$\begin{aligned}
\varepsilon^{4H}\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy\,r(x-y)^2&\le\varepsilon^{4H}\sigma_0^4\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy1_{|x-y|\le z_\eta}\\
&\quad+\varepsilon^{4H}\eta\int_0^{z/\varepsilon^2}dx\int_0^{z/\varepsilon^2}dy|r(x-y)|\\
&\le\varepsilon^{4H-2}C_\eta+\eta|z|^{2-\gamma}\tilde{C}_\eta\,.
\end{aligned}$$

Then

$$\limsup_{\varepsilon\to 0}\mathbb{E}\left[\left|v_1^\varepsilon(z)-\sqrt{\frac{c_\nu}{c_m}}w^\varepsilon(z)\right|^2\right]\le\eta|z|^{2-\gamma}\tilde{C}_\eta$$

for every $\eta>0$, which concludes the proof. □

We consider next v_2^ε and v_3^ε.

LEMMA 7. *For every $z\in[0,Z]$, as ε goes to 0, the finite-dimensional distributions of $v_2^\varepsilon(z)$ and $v_3^\varepsilon(z)$ converge to those of the 0 process.*

Proof. Without loss of generality we present the proof only for $v_2^\varepsilon(z)$ and with $2\omega = 1$. We have

$$\begin{aligned}\mathbb{E}[v_2^\varepsilon(z)^2] &= \frac{1}{\varepsilon^{2\gamma}} \int_0^z dx \int_0^z dy \cos\left(\frac{x}{\varepsilon^\tau}\right) \cos\left(\frac{y}{\varepsilon^\tau}\right) r_\nu \left(\frac{x-y}{\varepsilon^2}\right) \\ &= I_1^\varepsilon(z) + I_2^\varepsilon(z)\,,\end{aligned}$$

with

$$\begin{aligned}I_1^\varepsilon(z) &= \frac{c_\nu}{\varepsilon^{2\gamma}} \int_0^z dx \int_0^z dy \cos\left(\frac{x}{\varepsilon^\tau}\right) \cos\left(\frac{y}{\varepsilon^\tau}\right) \left|\frac{x-y}{\varepsilon^2}\right|^{-\gamma}\,, \\ I_2^\varepsilon(z) &= \frac{1}{\varepsilon^{2\gamma}} \int_0^z dx \int_0^z dy \cos\left(\frac{x}{\varepsilon^\tau}\right) \cos\left(\frac{y}{\varepsilon^\tau}\right) \left(r_\nu\left(\frac{x-y}{\varepsilon^2}\right) - c_\nu \left|\frac{x-y}{\varepsilon^2}\right|^{-\gamma}\right)\,.\end{aligned}$$

Let $\delta > 0$; because $r_\nu(u) \sim c_\nu u^{-\gamma}$ as $u \to \infty$, we have that for $u > z_\delta$ (with z_δ sufficiently large) $|r_\nu(u) - c_\nu u^{-\gamma}| \le \delta u^{-\gamma}$. We then obtain

$$\begin{aligned}|I_2^\varepsilon(z)| &\le \frac{\delta}{\varepsilon^{2\gamma}} \int_0^z dx \int_0^z dy \cos\left(\frac{x}{\varepsilon^\tau}\right) \cos\left(\frac{y}{\varepsilon^\tau}\right) \left|\frac{x-y}{\varepsilon^2}\right|^{-\gamma} \\ &\quad + C \int_0^z dx \int_0^z dy |x-y|^{-\gamma} 1_{|x-y| \le \varepsilon^2 z_\delta} \\ &\le \delta \int_0^z dx \int_0^z dy |x-y|^{-\gamma} + C \int_0^z dx \int_0^z dy |x-y|^{-\gamma} 1_{|x-y|\le \varepsilon^2 z_\delta}\,,\end{aligned}$$

so that

$$\limsup_{\varepsilon\to 0} |I_2^\varepsilon(z)| \le \delta \int_0^z dx \int_0^z dy |x-y|^{-\gamma}\,.$$

The inequality above is valid for every $\delta > 0$, and we conclude

$$\lim_{\varepsilon \to 0} I_2^\varepsilon(z) = 0\,.$$

To complete the study of v_2^ε it remains to deal with $I_1^\varepsilon(z)$. We have

$$\begin{aligned}I_1^\varepsilon(z) &= c_\nu \varepsilon^{4-2\gamma} \int_0^{z/\varepsilon^2} dx \int_0^{z/\varepsilon^2} dy \cos(\varepsilon^{2-\tau} x) \cos(\varepsilon^{2-\tau} y) |x-y|^{-\gamma} \\ &= \frac{c_\nu}{2}\Big(I_{1,1}^\varepsilon(z) + I_{1,2}^\varepsilon(z)\Big)\,,\end{aligned}$$

where

$$\begin{aligned}I_{1,1}^\varepsilon(z) &= \varepsilon^{4-2\gamma} \int_0^{z/\varepsilon^2} dx \int_0^{z/\varepsilon^2} dy \cos(\varepsilon^{2-\tau}(x-y))|x-y|^{-\gamma}\,, \\ I_{1,2}^\varepsilon(z) &= \varepsilon^{4-2\gamma} \int_0^{z/\varepsilon^2} dx \int_0^{z/\varepsilon^2} dy \cos(\varepsilon^{2-\tau}(x+y))|x-y|^{-\gamma}\,.\end{aligned}$$

Using integration by parts, we get

$$I_{1,1}^\varepsilon(z) = 2\left(\varepsilon^{4-2\gamma} \int_0^{z/\varepsilon^2} dx \left(\int_0^{(z-x)/\varepsilon^2} dy \cos(\varepsilon^{2-\tau} y)|y|^{-\gamma}\right)\right) = 2(J_1^\varepsilon(z) - J_2^\varepsilon(z))\,,$$

where

$$J_1^\varepsilon(z) = \varepsilon^{2-2\gamma} z \int_0^{z/\varepsilon^2} dy \cos(\varepsilon^{2-\tau} y) y^{-\gamma}\,,$$
$$J_2^\varepsilon(z) = \varepsilon^{4-2\gamma} \int_0^{z/\varepsilon^2} dy \cos(\varepsilon^{2-\tau} y) y^{1-\gamma}\,.$$

We make the substitution $y \to y/\varepsilon^{2-\tau}$ in $J_1^\varepsilon(z)$ to obtain

$$J_1^\varepsilon(z) = \varepsilon^{\tau(1-\gamma)} z \int_0^{z/\varepsilon^\tau} dy \cos(y) y^{-\gamma}$$

and so $J_1^\varepsilon(z) = O(\varepsilon^{\tau(1-\gamma)})$. We also make the substitution $y \to y/\varepsilon^{2-\tau}$ in $J_2^\varepsilon(z)$ to obtain

$$J_2^\varepsilon(z) = \varepsilon^{\tau(2-\gamma)} \int_0^{z/\varepsilon^\tau} dy \cos(y) y^{1-\gamma}\,.$$

Using integration by parts, we obtain

$$\begin{aligned} J_2^\varepsilon(z) = \varepsilon^{\tau(2-\gamma)} \sin\left(\frac{z}{\varepsilon^\tau}\right) \left(\frac{z}{\varepsilon^\tau}\right)^{1-\gamma} \\ - \varepsilon^{\tau(2-\gamma)}(1-\gamma) \int_0^{z/\varepsilon^\tau} dy \sin(y) y^{-\gamma}\,, \end{aligned}$$

and so $J_2^\varepsilon(z) = O(\varepsilon^\tau)$. Therefore, we conclude

$$\lim_{\varepsilon\to 0} I_{1,1}^\varepsilon(z) = 0\,.$$

Consider finally $I_{1,2}^\varepsilon(z)$. Letting $x-y \to x$ and $x+y \to y$, we get

$$I_{1,2}^\varepsilon(s,t) = \varepsilon^{4-2\gamma} \iint_{D_1^\varepsilon \cup D_2^\varepsilon \cup D_3^\varepsilon \cup D_4^\varepsilon} dx\, dy\, |x|^{-\gamma} \frac{\cos(\varepsilon^{2-\tau} y)}{2},$$

where

$$\begin{aligned} D_1^\varepsilon &= \{(x,y) \in [0, z/\varepsilon^2] \times [0, z/\varepsilon^2] : x \le y\}\,, \\ D_2^\varepsilon &= \{(x,y) \in [0, z/\varepsilon^2] \times [z/\varepsilon^2, 2z/\varepsilon^2] : y \le -x + 2z/\varepsilon^2\}\,, \\ D_3^\varepsilon &= \{(x,y) \in [-z/\varepsilon^2, 0] \times [z/\varepsilon^2, 2z/\varepsilon^2] : y \le x + 2z/\varepsilon^2\}\,, \\ D_4^\varepsilon &= \{(x,y) \in [-z/\varepsilon^2, 0] \times [0, z/\varepsilon^2] : -y \le x\}\,. \end{aligned}$$

Let us deal with the integral on D_1^ε:

$$\varepsilon^{4-2\gamma} \iint_{D_1^\varepsilon} dx\, dy\, |x|^{-\gamma} \frac{\cos(\varepsilon^{2-\tau} y)}{2} = (2(1-\gamma))^{-1} J_2^\varepsilon(z) = O(\varepsilon^\tau)\,.$$

The integrals on D_2^ε, D_3^ε, and D_4^ε can be analyzed in a similar way; therefore, $I_{1,2}^\varepsilon(s,t) \to 0$. This finally shows

$$\lim_{\varepsilon\to 0} I_1^\varepsilon(z) = 0\,,$$

and then

$$\lim_{\varepsilon\to 0}\mathbb{E}[v_2^\varepsilon(z)^2]=0,$$

which concludes the proof. □

Now we deal with a technical lemma regarding increments of $\mathbf{v}^\varepsilon$.

LEMMA 8. *There exists a constant $C>0$ such that for every z, ζ, and $\varepsilon>0$ we have*

$$\mathbb{E}[\|\mathbf{v}^\varepsilon(z)-\mathbf{v}^\varepsilon(\zeta)\|^2]\le C|z-\zeta|^{2H}.$$

Proof. For every $j=1,2,3$, using (3.6), (3.7), and (A.3), we have (taking $z>\zeta$)

$$\begin{aligned}\mathbb{E}[|v_j^\varepsilon(z)-v_j^\varepsilon(\zeta)|^2] &\le \frac{1}{\varepsilon^{2\gamma}}\int_\zeta^z dx\int_\zeta^z dy\left|\mathbb{E}\left[\nu\left(\frac{x}{\varepsilon^2}\right)\nu\left(\frac{y}{\varepsilon^2}\right)\right]\right|\\
&\le \frac{C}{\varepsilon^{2\gamma}}\int_\zeta^z dx\int_\zeta^z dy\sum_{j=1}^\infty\left(\frac{J(j)}{j!}\right)^2\left|\mathbb{E}\left[H_j\left(\frac{m}{\sigma_0}\left(\frac{x}{\varepsilon^2}\right)\right)H_j\left(\frac{m}{\sigma_0}\left(\frac{y}{\varepsilon^2}\right)\right)\right]\right|\\
&\le \frac{C}{\varepsilon^{2\gamma}}\int_\zeta^z dx\int_\zeta^z dy\sum_{j=1}^\infty\frac{J(j)^2}{j!\sigma_0^{2j}}\left|r_m\left(\frac{x-y}{\varepsilon^2}\right)\right|^j\\
&\le \frac{C}{\varepsilon^{2\gamma}}\int_\zeta^z dx\int_\zeta^z dy\sum_{j=1}^\infty\frac{J(j)^2}{j!\sigma_0^{2}}\left|r_m\left(\frac{x-y}{\varepsilon^2}\right)\right|\\
&\le \frac{C'}{\varepsilon^{2\gamma}}\int_\zeta^z dx\int_\zeta^z dy\left|\frac{x-y}{\varepsilon^2}\right|^{-\gamma}\\
&\le \frac{2C'}{(1-\gamma)(2-\gamma)}\left|z-\zeta\right|^{2-\gamma},\end{aligned}$$

which concludes the proof. □

Using the above lemmas, we can deduce the following lemma, which deals with identification of the limit.

LEMMA 9. *As ε goes to 0, $\mathbf{v}^\varepsilon$ converges to W_H defined in Lemma 5 in the space of continuous functions endowed with the uniform norm.*

Proof. Lemmas 6 and 7 give the convergence of finite-dimensional distributions of $\mathbf{v}^\varepsilon$ to those of W_H. Using then the Kolmogorov criterion [2], Lemma 8, and the fact that $2H>1$, we get the tightness of $(\mathbf{v}^\varepsilon)_\varepsilon$ in the space of continuous functions endowed with the uniform norm, which establishes the proof. □

Thanks to Lemma 9 we conclude the proof of Lemma 5 by establishing the tightness in a rough paths sense.

LEMMA 10. *The sequence $(\mathbf{v}^\varepsilon)_\varepsilon$ is tight in Ω_p for $p>1/H$.*

Proof of Lemmas 5 *and* 10. Let $q\in(1/H,p)$. In view of Lemmas 3 and 9 it is enough to prove

$$\lim_{A\to+\infty}\sup_{\varepsilon>0}\mathbb{P}[V_q(\mathbf{v}^\varepsilon)>A]=0. \tag{5.4}$$

Using Chebyshev's inequality, the fact that $q < 2$, Lemma 4, the Hölder inequality, and Lemma 9, we find

$$\begin{aligned}\mathbb{P}[V_q(\mathbf{v}^\varepsilon) > A] &\le \frac{1}{A^q}\mathbb{E}[V_q(\mathbf{v}^\varepsilon)^q]\\ &\le \frac{C}{A^q}\sum_{n=1}^{+\infty} n^C \sum_{k=1}^{2^n} \mathbb{E}[\|\mathbf{v}^\varepsilon(z_k^n) - \mathbf{v}^\varepsilon(z_{k-1}^n)\|^q]\\ &\le \frac{C}{A^q}\sum_{n=1}^{+\infty} n^C \sum_{k=1}^{2^n} \mathbb{E}[\|\mathbf{v}^\varepsilon(z_k^n) - \mathbf{v}^\varepsilon(z_{k-1}^n)\|^2]^{q/2}\\ &\le \frac{C'}{A^q}\sum_{n=1}^{+\infty} n^C \sum_{k=1}^{2^n} \left(\frac{1}{2^n}\right)^{qH}\\ &\le \frac{C'}{A^q}\sum_{n=1}^{+\infty} n^C \left(\frac{1}{2^n}\right)^{qH-1},\end{aligned}$$

and since $qH > 1$ we deduce (5.4). □

Finally, we can now derive the following lemma, which deals with the convergence of the propagator.

LEMMA 11. *Let $\{\omega_1, \dots, \omega_n\}$ be a collection of frequencies. Then, as ε goes to 0, the propagator vector $(P^\varepsilon_{\omega_1}, \dots, P^\varepsilon_{\omega_n})$ converges to $(P_{\omega_1}, \dots, P_{\omega_n})$, which is the asymptotic propagator P_ω that we can write as*

$$P_\omega(z) = \begin{pmatrix} \exp\left(i\omega c_H/2B_H(z)\right) & 0 \\ 0 & \exp\left(-i\omega c_H/2B_H(z)\right)\end{pmatrix}.$$

Proof. By combining Theorem 2, (5.1), and Lemma 5, we get that, as ε goes to 0, P^ε_ω converges in distribution in the space of continuous functions (endowed with the uniform topology) to the solution P_ω of the following system of equations in the sense of rough paths:

$$dP_\omega(z) = \frac{i\omega c_H}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1\end{pmatrix} P_\omega(z)\, dB_H(z)\,.$$

This concludes the proof. □

5.3. Conclusion of the proof. The remaining part of the proof of Theorem 1 follows the lines of [6, 14]; however, we present it here for completeness. Recall that, thanks to the formula (2.12), we can write $a^\varepsilon(Z, s)$ in a Fourier-type formula using the transmission coefficient,

$$a^\varepsilon(Z, s) = \frac{1}{2\pi}\int e^{-is\omega} T^\varepsilon_\omega(Z)\widehat{f}(\omega)\, d\omega\,, \tag{5.5}$$

with the transmission coefficient being a functional of the propagator P^ε_ω. We shall use Lemma 11 to deduce the convergence of the transmitted wave.

Let $n \in \mathbb{N}$, $s_1 \le \cdots \le s_n \in [0, \infty)$. We can write

$$\begin{aligned}\mathbb{E}[a^\varepsilon(Z, s_1)\cdots a^\varepsilon(Z, s_n)] &= \mathbb{E}\left[\frac{1}{(2\pi)^n}\prod_{j=1}^n \int e^{-is_j\omega} T^\varepsilon_\omega(Z)\widehat{f}(\omega)\, d\omega\right]\\ &= \frac{1}{(2\pi)^n}\int\cdots\int e^{-i\sum_{j=1}^n s_j\omega_j}\widehat{f}(\omega_1)\cdots\widehat{f}(\omega_n)\mathbb{E}[T^\varepsilon_{\omega_1}(Z)\cdots T^\varepsilon_{\omega_n}(Z)]\, d\omega_1\cdots d\omega_n.\end{aligned}$$

Thanks to Lemma 11 we have that as $\varepsilon \to 0$,

$$\mathbb{E}[T^\varepsilon_{\omega_1}(Z)\cdots T^\varepsilon_{\omega_n}(Z)] \to \mathbb{E}\left[\exp\left(\frac{ic_H B_H(Z)}{2}\sum_{j=1}^n \omega_j\right)\right],$$

and then

$$\begin{aligned}\mathbb{E}[a^\varepsilon(Z,s_1)\cdots a^\varepsilon(Z,s_n)] &\to \frac{1}{(2\pi)^n}\int\cdots\int e^{-i\sum_{j=1}^n s_j\omega_j}\widehat{f}(\omega_1)\cdots\widehat{f}(\omega_n)\\ &\quad\times\mathbb{E}\left[\exp\left(\frac{ic_H B_H(Z)}{2}\sum_{j=1}^n \omega_j\right)\right] d\omega_1\cdots d\omega_n\\ &= \mathbb{E}\left[\frac{1}{(2\pi)^n}\prod_{j=1}^n\int e^{-i(s_j - c_H B_H(Z)/2)\omega}\widehat{f}(\omega)\,d\omega\right]\\ &= \mathbb{E}\left[\prod_{j=1}^n f\left(s_j - \frac{c_H B_H(Z)}{2}\right)\right].\end{aligned}$$

The tightness proof is similar to the proof of Lemma 3.2 in [6], and the convergence of $a^\varepsilon(Z,s)$ follows. To conclude the proof of Theorem 1 it remains to prove the convergence of $b^\varepsilon(0,s)$. It is similar to the convergence of $a^\varepsilon(Z,s)$, up to substituting the application of (2.12) by that of (2.13).

Appendix A. Hermite polynomials.

In this appendix we recall some results regarding Hermite polynomials that we use in this paper. We denote the Gaussian probability density of a random variable $X \sim \mathcal{N}(0,1)$ by

$$g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}},$$

and we define for every $k \in \mathbb{N}$ the kth Hermite polynomial by

$$H_k(x) = (-1)^k\frac{g^{(k)}(x)}{g(x)}.$$

The set of all Hermite polynomials $\{H_k,\ k = 0,1,2,\dots\}$ is an orthonormal base for the space $L^2(g(x)\,dx) = \{h : \mathbb{E}[|h(X)|^2] < \infty\}$. We denote by $J_h(k)$ (or $J(k)$ if there is no ambiguity) the projection coefficient of a function $h \in L^2(g(x)\,dx)$ on the subspace spanned by H_k, that is,

$$J_h(k) = \mathbb{E}[H_k(X)h(X)].$$

Then, we have the series representation

$$h(x) = \sum_{k=0}^\infty \frac{J_h(k)}{k!}H_k(x), \tag{A.1}$$

the convergence being in $L^2(g(x)\,dx)$. We can explicitly compute the second moments by

$$\mathbb{E}[|h(X)|^2] = \sum_{k=0}^\infty \frac{|J_h(k)|^2}{k!}. \tag{A.2}$$

This formula is a direct consequence of the following relation that we use in this paper: for a centered two-dimensional Gaussian vector (X_1, X_2) such that $\mathbb{E}[X_1^2] = \mathbb{E}[X_2^2] = 1$ we have

$$\mathbb{E}[H_j(X_1)H_k(X_2)] = \begin{cases} k!\mathbb{E}[X_1X_2]^k & \text{if} \quad k = l, \\ 0 & \text{if} \quad k \neq l. \end{cases} \tag{A.3}$$

REFERENCES

[1] M. Asch, W. Kohler, G. C. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[2] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[3] L. Borcea, G. Papanicolaou, and C. Tsogka, *Theory and applications of time reversal and interferometric imaging*, Inverse Problems, 19 (2003), pp. S134–S164.

[4] L. Borcea, G. Papanicolaou, and C. Tsogka, *Coherent interferometry in finely layered random media*, Multiscale Model. Simul., 5 (2006), pp. 62–83.

[5] S. A. Bulgakov, V. V. Konotop, and L. Vazquez, *Wave interaction with a random fat fractal: Dimension of the reflection coefficient*, Waves in Random Media, 5 (1995), pp. 9–18.

[6] J. F. Clouet and J. P. Fouque, *Spreading of a pulse travelling in random media*, Ann. Appl. Probab., 4 (1994), pp. 1083–1097.

[7] L. Coutin, *An introduction to (stochastic) calculus with respect to fractional Brownian motion*, in Séminaire de Probabilités XL, Lecture Notes in Math. 1899, Springer, New York, 2007, pp. 3–65.

[8] L. Coutin and Z. Qian, *Stochastic analysis, rough path analysis and fractional Brownian motions*, Probab. Theory Related Fields, 122 (2002), pp. 108–140.

[9] S. Dolan, C. Bean, and B. Riollet, *The broad-band fractal nature of heterogeneity in the upper crust from petrophysical logs*, Geophys. J. Int., 132 (1998), pp. 489–507.

[10] A. Fannjiang and K. Solna, *Scaling limits for laser beam propagation in atmospheric turbulence*, Stoch. Dyn., 4 (2004), pp. 135–150.

[11] A. C. Fannjiang and K. Solna, *Propagation and time reversal of wave beams in atmospheric turbulence*, Multiscale Model. Simul., 3 (2005), pp. 522–558.

[12] A. Fannjiang and K. Solna, *Superresolution and duality for time-reversal of waves in random media*, Phys. Lett. A, 352 (2005), pp. 22–29.

[13] J. Feder, *Fractals*, Plenum Press, New York, 1988.

[14] J. P. Fouque, J. Garnier, G. Papanicolaou, and K. Solna, *Wave Propagation and Time Reversal in Randomly Layered Media*, Springer, New York, 2007.

[15] S. Frey, W. Geurts, and L. Woste, *Laser remote sensing for characterization of planetary boundary layer properties*, in CLEO 2001, Technical Digest, Optical Society of America, Washington, DC, 2001, p. 494.

[16] A. E. Gargett, *The scaling of turbulence in the presence of stable stratification*, J. Geophys. Res., 93 (1988), pp. 5021–5036.

[17] J. Garnier and K. Sølna, *Effective transport equations and enhanced backscattering in random waveguides*, SIAM J. Appl. Math., 68 (2008), pp. 1574–1599.

[18] J. Garnier and K. Solna, *Random backscattering in the parabolic scaling*, J. Statist. Phys., 131 (2008), pp. 445–486.

[19] F. Herrmann, *A Scaling Medium Representation, A Discussion on Well-Logs, Fractals and Waves*, Ph.D. thesis, Department of Geophysics, Delft University of Technology, Delft, The Netherlands, 1997.

[20] T. A. Hewett, *Modeling reservoir heterogeneities with fractals*, in Proceedings of the 4th International Geostatistics Congress, Terra Abstracts 4, Suppl. 3, 9, 1992.

[21] V. V. Konotop, Z. Fei, and L. Vazquez, *Wave interaction with a fractal layer*, Phys. Rev. E, 48 (1993), pp. 4044–4048.

[22] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1994.

[23] M. Ledoux, T. Lyons, and Z. Qian, *Lévy area of Wiener processes in Banach spaces*, Ann. Probab., 30 (2002), pp. 546–578.

[24] A. Lejay, *An introduction to rough paths*, in Séminaire de Probabilités XXXVII, Lecture Notes in Math., Springer-Verlag, New York, Berlin, 2003.

[25] P. Lewicki, R. Burridge, and M. V. de Hoop, *Beyond effective medium theory: Pulse stabilization for multimode wave propagation in high-contrast layered media*, SIAM J. Appl. Math., 56 (1996), pp. 256–276.

[26] P. Lewicki, R. Burridge, and G. Papanicolaou, *Pulse stabilization in a strongly heterogeneous medium*, Wave Motion, 20 (1994), pp. 177–195.

[27] T. Lyons, *Differential equations driven by rough signals*, Rev. Mat. Iberoamer., 14 (1998), pp. 215–310.

[28] T. Lyons, *Differential equations driven by rough signals* (I): *An extension of an inequality of L. C. Young*, Math. Res. Lett., 1 (1994), pp. 451–464.

[29] T. Lyons and Z. Qian, *System Control and Rough Paths*, Oxford Mathematical Monographs, Oxford University Press, London, 2002.

[30] R. Marty, *Théorème limite pour une équation différentielle à coefficient aléatoire à mémoire longue*, C. R. Acad. Sci. Paris Ser. I, 338 (2004), pp. 167–170.

[31] R. Marty, *Asymptotic behavior of differential equations driven by periodic and random processes with slowly decaying correlations*, ESAIM Probab. Statist., 9 (2005), pp. 165–184.

[32] R. F. O'Doherty and N. A. Anstey, *Reflections on amplitudes*, Geophys. Prospecting, 19 (1971), pp. 430–458.

[33] H. Omre, K. Solna, and H. Tjelmeland, *Simulation of random functions on large lattices*, in Geostatistics, A. Soares, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 179–199.

[34] M. Pilkington and J. P. Todoeschuck, *Stochastic inversion for scaling geology*, Geophys. J. Int., 102 (1990), pp. 205–217.

[35] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman and Hall, London, 1994.

[36] C. Sidi and F. Dalaudier, *Turbulence in the stratified atmosphere: Recent theoretical developments and experimental results*, Adv. Space Res., 10 (1990), pp. 25–36.

[37] K. Sølna, *Acoustic pulse spreading in a random fractal*, SIAM J. Appl. Math., 63 (2003), pp. 1764–1788.

[38] K. Solna and G. Papanicolaou, *Ray theory for a locally layered medium*, Waves in Random Media, 10 (2000), pp. 151–198.

[39] X. Sun and D. L. Jaggard, *Wave interaction with a generalized Cantor bar fractal multilayers*, J. Appl. Phys., 70 (1991), pp. 2500–2507.

[40] M. S. Taqqu, *Weak convergence to fractional Brownian motion and to the Rosenblatt process*, Z. Wahrsch. Verw. Gebiete, 31 (1975), pp. 287–302.

[41] M. S. Taqqu, *Convergence of integrated processes of arbitrary Hermite rank*, Z. Wahrsch. Verw. Gebiete, 50 (1979), pp. 53–83.

[42] E. Tromeur, E. Garnier, P. Sagaut, and C. Basdevant, *Large eddy simulations of aero-optical effects in a turbulent boundary layer*, J. Turbulence, 4 (2003), pp. 1–22.

[43] D. Washburn, D. W. Banton, T. T. Brennan, W. P. Brown, R. R. Butts, S. C. Coy, R. H. Dueck, K. W. Koenig, B. S. Masson, P. H. Peterson, R. W. Praus, G. A. Tyler, B. P. Venet, and L. D. Weaver, *Airborne Laser Extended Atmospheric Characterization Experiment (ABLE ACE)*, technical report, Phillips Laboratory, Kirtland Air Force Base, Albuquerque, NM, 1996.

[44] J. Weiss and D. Marsan, *Scale properties of sea ice deformation and fracturing*, C. R. Phys., 5 (2004), pp. 736–751.

SIAM J. APPL. MATH.
Vol. 69, No. 4, pp. 1084–1109

THE UNSTEADY FLOW OF A WEAKLY COMPRESSIBLE FLUID IN A THIN POROUS LAYER I: TWO-DIMENSIONAL THEORY[*]

D. J. NEEDHAM[†], S. LANGDON[‡], G. S. BUSSWELL[§], AND J. P. GILCHRIST[§]

Abstract. We consider the problem of determining the pressure and velocity fields for a weakly compressible fluid flowing in a two-dimensional reservoir in an inhomogeneous, anisotropic porous medium, with vertical side walls and variable upper and lower boundaries, in the presence of vertical wells injecting or extracting fluid. Numerical solution of this problem may be expensive, particularly in the case that the depth scale of the layer h is small compared to the horizontal length scale l. This is a situation which occurs frequently in the application to oil reservoir recovery. Under the assumption that $\epsilon = h/l \ll 1$, we show that the pressure field varies only in the horizontal direction away from the wells (the outer region). We construct two-term asymptotic expansions in ϵ in both the inner (near the wells) and outer regions and use the asymptotic matching principle to derive analytical expressions for all significant process quantities. This approach, via the method of matched asymptotic expansions, takes advantage of the small aspect ratio of the reservoir, ϵ, at precisely the stage where full numerical computations become stiff, and also reveals the detailed structure of the dynamics of the flow, both in the neighborhood of wells and away from wells.

Key words. oil recovery, thin porous layer, matched asymptotics

AMS subject classifications. 35K15, 35K20, 76M45, 76S05, 86A99

DOI. 10.1137/070703405

1. Introduction.

It is standard practice in the oil and gas industry to use reservoir simulators based on numerical methods such as the finite difference or finite element techniques. This kind of approach has been shown to be enormously successful over the years in modeling a wide variety of physical processes in the reservoir e.g., faults, rock layering effects, complex fluid phase behavior, etc. While reservoir simulators of this type will continue to play a crucial role in the industry, it is well known that to use them takes considerable expertise and time. Because of the numerical nature of the modeling process, gridding, time-stepping, and convergence issues require care and attention. Long execution times are often necessary for certain types of problems, e.g., hydraulically fractured wells.

Analytic techniques, for the reasons outlined, can therefore play a valuable role in the industry. Such techniques, although they may have some simplifying assumptions, allow a reservoir or production engineer to perform a quick study of their reservoir in order to obtain a broad understanding of the dynamical processes and make approximate costing forecasts. Analytic solutions are extremely fast and provide none of the timestepping and convergence issues seen with a numerically based simulator. Also, a necessary step in many reservoir studies involves the history matching of observed data by optimizing model parameters. The history matched model is then used for

[*] Received by the editors September 21, 2007; accepted for publication (in revised form) October 6, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/siap/69-4/70340.html

[†] School of Mathematics, University of Birmingham, Birmingham B15 2TT, UK (d.j.needham@bham.ac.uk).

[‡] Department of Mathematics, University of Reading, Reading RG6 6AX, UK (s.langdon@reading.ac.uk).

[§] Schlumberger Oilfield UK Plc, Wyndyke Furlong, Abingdon, OX14 1UJ, UK (gbusswell@hotmail.co.uk, pgilchrist@abingdon.oilfield.slb.com).

performance prediction. Given the speed and reliability of analytic results, there is a clear opportunity to exploit their use in history matching studies.

There has been much work in the literature regarding analytic approaches, particularly for well testing applications [10, 5, 15, 1, 9] but also from a full field reservoir standpoint, where multiple wells and reservoir boundaries must be accounted for to forecast production over the required timescales. Algorithms for full field simulation problems based on analytic approaches have been presented in the literature for porous media with homogeneous and anisotropic permeability in a variety of sources [4, 19, 13]. A more complex problem involves the application of analytic approaches to full field scenarios where the reservoir has inhomogeneous permeability and variable geometry [12, 14, 8, 18].

In this paper we introduce a new approach to solving full field reservoir problems with inhomogeneous and anisotropic permeability and variable reservoir geometry using the method of matched asymptotic expansions. Specifically, our problem involves determining analytical expressions for the pressure and velocity fields for a weakly compressible fluid flowing in a horizontal reservoir with variable upper and lower boundaries. Vertical wells injecting or extracting fluid from the reservoir can be considered as line sources and sinks, respectively. Numerical solution of the full equations of motion throughout the reservoir can be prohibitively expensive. However, under the condition that the depth scale of the reservoir h is small compared to the length scale of the reservoir l, as is often the case in geophysical applications, it can be shown (allowing further that the porous medium has inhomogeneous and anisotropic permeability) that the dimension of the problem can be reduced away from the wells, with solution of the full equations of motion being required only in a small domain around the wells where the geometry is radically simplified. Moreover, as the ratio h/l decreases, efficient application of numerical schemes becomes harder, while the problem becomes more amenable to solution via matched asymptotic theory.

Here, we restrict attention to the case of two-dimensional flow. The full three-dimensional problem will be dealt with in subsequent work [11]. We introduce the parameter $\epsilon = h/l$ and consider asymptotic solutions to the equations of motion of the fluid in increasing powers of ϵ, with $0 < \epsilon \ll 1$. In the vicinity of a well (the *inner* region) the pressure field is two-dimensional, but away from the wells (the *outer* region) the pressure field is only one-dimensional. This immediately leads to a reduction in complexity. Here, however, rather than solving the full equations of motion numerically in the inner and outer regions, we construct two-term expansions in both the inner and outer regions. These expansions in the inner and outer regions can then be matched, via the Van Dyke asymptotic matching principle [20], enabling us to derive amenable analytical expressions for all significant process quantities.

We begin in section 2 by deriving the equations of motion in the porous medium. Conservation of mass and momentum lead to a strongly parabolic linear initial boundary value problem for the dynamic fluid pressure (from which the fluid velocity field can be deduced), with Neumann boundary conditions, under the assumption that the walls are impenetrable to the fluid in the porous medium. This initial boundary value problem has a unique solution, but its direct computation would be expensive, primarily due to stiffness when $0 < \epsilon \ll 1$. We thus consider the associated steady state problem [SSP], a linear strongly elliptic Neumann problem, which also has a unique solution (up to a constant) under the further constraint that the sum of the total volume fluxes at the wells (the line sources and sinks) is zero. Solution of the steady state problem is considered in section 3. Subtracting the solution of the steady state problem from the solution of the initial value problem leads to a strongly parabolic homogeneous problem with no discontinuities across the sources and sinks. The solution of this problem leads to a regular self-adjoint eigenvalue problem [EVP] whose solution is considered in section 4.

Rather than solving [SSP] and [EVP] directly, the solution to each problem is considered in the asymptotic limit $\epsilon \to 0$, via the method of matched asymptotic expansions. For the two-dimensional problem these asymptotic solutions can be constructed analytically. To solve [SSP], we proceed first with the situation when the wells are well spaced and are away from the reservoir boundaries, after which the case of wells close to a boundary, or close together, is considered in sections 3.1 and 3.2. The asymptotic solution can be constructed directly in the outer region, up to $O(\epsilon^2)$. In the inner region, determination of the leading order terms reduces to the solution of a strongly elliptic problem whose solution can be written analytically in terms of the eigenvalues and corresponding eigenfunctions of a regular Sturm–Liouville eigenvalue problem. The asymptotic solution of [EVP] in section 4 also reduces to a regular Sturm–Liouville eigenvalue problem identical in structure to that discussed in section 3, and a consideration of this allows us to demonstrate that the solution to the full initial boundary value problem approaches the solution to the steady state problem through terms exponentially small with respect to time t as $t \to \infty$. With D_z being the permeability scale in the vertical direction and D_x being the permeability scale in the horizontal direction, the further generalization that $D_z = o(D_x)$ rather than $O(D_x)$ is considered in section 5, where it is shown that the structure of the solution is identical to that found for the case that $D_z = O(D_x)$, after a suitable redefinition of the parameter ϵ. The constraint on the sum of the total volume fluxes at the wells being zero is removed in section 6, and in section 7 we apply the theory to a simple model example. Finally in section 8 we draw some conclusions.

2. Equations of motion.

We consider the flow of a weakly compressible fluid in the presence of sources and sinks in a reservoir of porous medium with variable upper and lower boundary. The reservoir has permeability which is inhomogeneous and anisotropic. We restrict attention to the situation when the flow is two-dimensional. We denote the interior of the porous medium by $M \subset \mathbb{R}^2$ and its impermeable boundary by $\partial M \subset \mathbb{R}^2$, with $\bar{M} = M \cup \partial M$. We introduce rectangular Cartesian coordinates (x, z), with z pointing vertically upwards and x pointing horizontally. The vertical side walls of the reservoir are taken to be at $x = \pm l$, with $l > 0$. The upper and lower surface of the reservoir are described by $z = hz_+(x/l)$ and $z = hz_-(x/l)$, respectively, for $x \in [-l, l]$, with h (> 0) being the reservoir depth scale and $z_+, z_- : [-1, 1] \mapsto \mathbb{R}$ being such that $z_+, z_- \in C^1([-1, 1])$ and $z_+(x) > z_-(x)$ for all $x \in [-1, 1]$. Normal fields on the upper and lower surfaces are then given by $\mathbf{n}_+(x) = (-\frac{h}{l}z'_+(x), 1)$ and $\mathbf{n}_-(x) = (\frac{h}{l}z'_-(x), -1)$, respectively, for $x \in [-1, 1]$, with the normals directed out of $\bar{M}$. The situation is illustrated in Figure 2.1.

Embedded within $\bar{M}$ are $N (\in \mathbb{N})$ vertical line sources/sinks at locations $x_i \in (-l, l)$, $i = 1, \ldots, N$. Each line source/sink extends from the upper surface to the lower surface of $\bar{M}$ and represents a model of a vertical bore hole in the reservoir extending from the upper to the lower surface of the reservoir and extracting or injecting fluid along its whole length. This model is standard in the oil industry [3, 10]. The prescribed strength of each line source/sink then represents the details of the controlled volumetric extraction mechanism in the bore hole along its length. The two components of permeability, in the x- and z-directions, respectively, are given by

$$D_0 D_x\left(\frac{x}{l}, \frac{z}{h}\right) \geq D_m > 0, \quad D_0 D_z\left(\frac{x}{l}, \frac{z}{h}\right) \geq D_m > 0, \quad (x, z) \in \bar{M}, \tag{2.1}$$

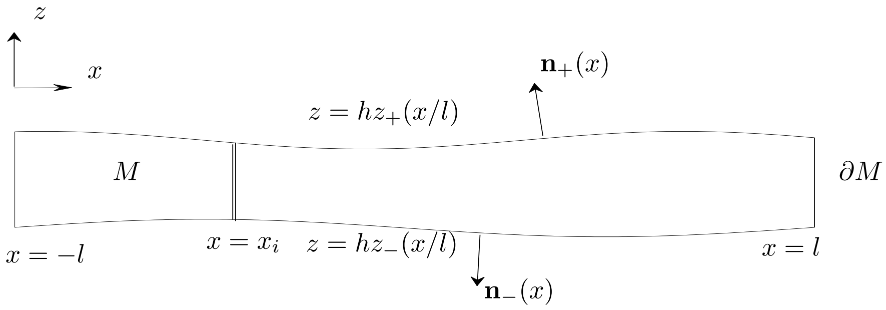


FIG. 2.1. *Porous layer $M \subset \mathbb{R}^2$, with impermeable boundary ∂M.*

with $D_x, D_z : \bar{M} \mapsto \mathbb{R}^+$ such that

$$D_x, D_z \in C^1(\bar{M}). \tag{2.2}$$

Here $D_0 > 0$ is a permeability scale for the layer and $D_m > 0$ is a lower bound on permeability in the layer. To avoid confusion, we emphasise here that subscripts x and z attached to the functions D_x and D_z do not indicate partial differentiation but merely denote the permeability direction.

We represent the fluid velocity field and pressure field at each point within the layer by $\mathbf{q} = \mathbf{q}(\mathbf{r}, t) = (u(\mathbf{r}, t), w(\mathbf{r}, t))$ and $p = p(\mathbf{r}, t)$, respectively, for each $(\mathbf{r}, t) \in \bar{M} \times [0, \infty)$. Here $t \geq 0$ represents time. The equation of conservation of fluid mass within the layer is then

$$\rho_t + (\rho u)_x + (\rho w)_z = \sum_{i=1}^{N} \rho s_i \left(\frac{z}{h}\right) \frac{1}{l} \delta \left(\frac{x - x_i}{l}\right), \quad (x, z) \in M, \quad t \in (0, \infty). \tag{2.3}$$

Here $\delta : \mathbb{R} \mapsto \mathbb{R}$ is the usual Dirac delta function, $\rho = \rho(\mathbf{r}, t)$ is the fluid density field for $(\mathbf{r}, t) \in \bar{M} \times [0, \infty)$, and the line source/sink volumetric strengths are represented by $s_i : \left[z_-\left(\frac{x_i}{l}\right), z_+\left(\frac{x_i}{l}\right)\right] \mapsto \mathbb{R}$, $i = 1, \ldots, N$. From practical considerations of the employed volumetric extraction mechanisms in bore holes, it is reasonable to take

$$s_i \in C\left(\left[z_-\left(\frac{x_i}{l}\right), z_+\left(\frac{x_i}{l}\right)\right]\right), \quad i = 1, \ldots, N. \tag{2.4}$$

The total volume flux per unit width from the ith line source/sink is then

$$Q_i = \int_{hz_-(x_i/l)}^{hz_+(x_i/l)} s_i \left(\frac{\lambda}{h}\right) \mathrm{d}\lambda, \quad i = 1, \ldots, N. \tag{2.5}$$

Conservation of momentum in the fluid is accounted for through the D'Arcy equation for flow in a porous media, giving

$$u = -D_0 D_x \left(\frac{x}{l}, \frac{z}{h}\right) p_x, \quad w = -D_0 D_z \left(\frac{x}{l}, \frac{z}{h}\right) (p_z + \rho g) \tag{2.6}$$

for all $(x, z) \in M$, $t \in (0, \infty)$, where g is the acceleration due to gravity. The effect of weak compressibility is accounted for through the equation of state,

$$\rho(p) = \rho_0(1 + c_t(p - p_0)), \tag{2.7}$$

with $c_t > 0$ being a constant isothermal expansion coefficient and ρ_0 and p_0 being positive reference density and pressure, respectively. Now, for a weakly compressible fluid, $0 < c_t p_0 \ll 1$, and the weakly compressible model is obtained by replacing $\rho(p)$ by its dominant contribution from (2.7) in both of the equations of motion (2.3) and (2.6). We obtain as our final model for the flow in the reservoir

$$c_t p_t + u_x + w_z = \sum_{i=1}^{N} s_i\left(\frac{z}{h}\right)\frac{1}{l}\delta\left(\frac{x-x_i}{l}\right), \tag{2.8}$$

$$u = -D_0 D_x\left(\frac{x}{l},\frac{z}{h}\right)p_x, \tag{2.9}$$

$$w = -D_0 D_z\left(\frac{x}{l},\frac{z}{h}\right)(p_z + \rho_0 g) \tag{2.10}$$

for all $(x, z) \in M$, $t \in (0, \infty)$. The equations of motion (2.8)–(2.10) form the basis for established models for weakly compressible flows in porous reservoirs [2], and we will take (2.8)–(2.10) as the model for the flow in the porous reservoir throughout the rest of the paper. We now set

$$Q = \sum_{i=1}^{N} |Q_i| \quad (> 0). \tag{2.11}$$

The natural scales are then $x \sim l$ and $z \sim h$, from the geometry of the porous layer, while $s_i \sim Q/h$, via (2.5). The continuity equation (2.8) then requires $u \sim Q/h$ and $w \sim Q/l$, while the momentum equation (2.9) requires $p \sim (lQ)/(hD_0)$. We therefore introduce the dimensionless variables,

$$x = lx',\ z = hz',\ s_i = \frac{Q}{h}s_i',\ u = \frac{Q}{h}u',\ w = \frac{Q}{l}w',\ p = \left(\frac{lQ}{hD_0}\right)p',\ t = \frac{c_t l^2}{D_0}t'. \tag{2.12}$$

On substitution from (2.12) into (2.8)–(2.10) (and dropping primes for convenience) we obtain the dimensionless equations of motion as

$$\bar{p}_t + u_x + w_z = \sum_{i=1}^{N} s_i(z)\delta(x - x_i), \tag{2.13}$$

$$u = -D_x(x,z)\bar{p}_x, \tag{2.14}$$

$$\epsilon^2 w = -D_z(x,z)\bar{p}_z \tag{2.15}$$

for all $(x, z) \in M'$, $t \in (0, \infty)$. Here, $p(x, z, t) = -\hat{\sigma}z + \bar{p}(x, z, t)$, with $\bar{p}$ being the dynamic fluid pressure, and the dimensionless parameters ϵ and $\hat{\sigma}$ are given by $\epsilon = h/l$ and $\hat{\sigma} = h^2\rho_0 g D_0/(lQ)$. The values of the model parameters will vary depending upon the details of the reservoir under consideration. However, for a typical field scenario the values $h \sim 200\,\mathrm{m}$, $l \sim 20{,}000\,\mathrm{m}$, $Q \sim 2\,\mathrm{m}^3/\mathrm{s}$, $c_t \sim 1.45 \times 10^{-9}\,\mathrm{m}^2/\mathrm{N}$, $p_0 \sim 2.76 \times 10^7\,\mathrm{N/m}^2$, and $D_0 \sim 10^{-10}\,\mathrm{m}^4/\mathrm{Ns}$ may be considered as representative. This gives a value for $c_t p_0 \sim 0.04$, which is entirely consistent with the adoption of the weakly compressible model proposed earlier. More significantly, the aspect ratio of a typical reservoir gives $\epsilon \sim 0.01$ (and this may be as small as 10^{-3} for large reservoirs). The dimensionless domain is now

$$M' = \{(x, z) \in \mathbb{R}^2 : x \in (-1, 1),\ z \in (z_-(x), z_+(x))\},$$

with closure $\bar{M}'$ and boundary $\partial M'$. The line source/sink locations are at $x_i \in (-1,1)$, $i = 1, \dots, N$. The volume flux condition (2.5) becomes

$$\alpha_i = \int_{z_-(x_i)}^{z_+(x_i)} s_i(\mu)\,\mathrm{d}\mu, \quad i = 1, \dots, N,$$

with $\alpha_i = Q_i/Q$, and hence $|\alpha_i| = |Q_i|/Q \le 1$, $i = 1, \dots, N$, and $\sum_{i=1}^N |\alpha_i| = \sum_{i=1}^N |Q_i|/Q = 1$, via (2.11). We next consider the boundary conditions. On the boundary $\partial M'$ the walls are impenetrable to the fluid in the porous layer. Thus

$$u(-1,z,t) = 0 \quad \text{for all } z \in [z_-(-1), z_+(-1)],\ t \in (0,\infty), \tag{2.16}$$

$$u(1,z,t) = 0 \quad \text{for all } z \in [z_-(1), z_+(1)],\ t \in (0,\infty), \tag{2.17}$$

$$w(x, z_+(x), t) - z_+'(x) u(x, z_+(x), t) = 0 \quad \text{for all } x \in (-1,1),\ t \in (0,\infty), \tag{2.18}$$

$$w(x, z_-(x), t) - z_-'(x) u(x, z_-(x), t) = 0 \quad \text{for all } x \in (-1,1),\ t \in (0,\infty). \tag{2.19}$$

Finally we have the initial condition

$$\bar{p}(x,z,0) = \bar{p}_0(x,z) \quad \text{for all } (x,z) \in \bar{M}', \tag{2.20}$$

with $\bar{p}_0 \in C(\bar{M}') \cap PC^1(\bar{M}')$, where $PC^1(\bar{M}')$ represents the class of piecewise continuously differentiable functions on $\bar{M}'$. The full problem for consideration is given by (2.13)–(2.15), (2.16)–(2.20), which we will refer to as [IBVP]. To proceed it is convenient to introduce $d_i = \{(x,z) \in \bar{M}' : x = x_i,\ z \in (z_-(x_i), z_+(x_i))\} \subset \bar{M}'$, for each $i = 1, \dots, N$, and $d = \bigcup_{i=1}^N d_i$. We require that a solution to [IBVP] has the following regularity, which is classical in the framework of the Dirac delta function formalism:

(i) $\bar{p} \in C(\bar{M}' \times [0,\infty)) \cap C^1((\bar{M}' \backslash \bar{d}) \times (0,\infty)) \cap C^2((M' \backslash d) \times (0,\infty))$,
$u \in C((\bar{M}' \backslash \bar{d}) \times (0,\infty)) \cap C^1((M' \backslash d) \times (0,\infty))$,
$w \in C(\bar{M}' \times (0,\infty)) \cap C^1((M' \backslash d) \times (0,\infty))$;

(ii) $\lim_{x \to x_i^\pm} \bar{p}_x$ and $\lim_{x \to x_i^\pm} \bar{p}_z$ both exist uniformly for $z \in [z_-(x_i), z_+(x_i)]$ at each $t \in (0,\infty)$, $i = 1, \dots, N$;

(iii) $[\bar{p}_z]_{x_i^-}^{x_i^+} = 0$, $[\bar{p}_x]_{x_i^-}^{x_i^+} = -s_i(z)/D_x(x_i, z)$ for all $z \in [z_-(x_i), z_+(x_i)]$ at each $t \in (0,\infty)$, $i = 1, \dots, N$.

We observe that (ii) requires, via (2.14), (2.15), that $\lim_{x \to x_i^\pm} u$ and $\lim_{x \to x_i^\pm} w$ both exist uniformly for $z \in [z_-(x_i), z_+(x_i)]$, and (iii) requires that

$$[u]_{x_i^-}^{x_i^+} = s_i(z), \quad [w]_{x_i^-}^{x_i^+} = 0$$

for all $z \in [z_-(x_i), z_+(x_i)]$ at each $t \in (0,\infty)$, $i = 1, \dots, N$. We now have the following preliminary result concerning [IBVP].

THEOREM 2.1. *For each $\epsilon > 0$,* [IBVP] *has a unique solution $u, w, \bar{p} : \bar{M}' \times [0,\infty) \mapsto \mathbb{R}$. Moreover,*

$$\left(\iint_{\bar{M}'} \bar{p}(x,z,t)\,\mathrm{d}x\,\mathrm{d}z \right)_t = \sum_{i=1}^N \alpha_i \tag{2.21}$$

for all $t \in (0,\infty)$.

Proof. Existence and uniqueness follow via regularity (2.2) and (2.4), with (2.1), on noting from (2.13)–(2.15) that [IBVP] is equivalent to the scalar initial boundary value problem

$$\bar{p}_t - \left\{(D_x(x,z)\bar{p}_x)_x + \left(\epsilon^{-2}D_z(x,z)\bar{p}_z\right)_z\right\} = \sum_{i=1}^{N} s_i(z)\delta(x-x_i), \tag{2.22}$$
$$(x,z)\in M', \quad t\in(0,\infty),$$
$$\bar{p}_x(-1,z,t)=0 \text{ for all } z\in[z_-(-1),z_+(-1)],\ t\in(0,\infty),$$
$$\bar{p}_x(1,z,t)=0 \text{ for all } z\in[z_-(1),z_+(1)],\ t\in(0,\infty),$$
$$D_z(x,z_+(x))\bar{p}_z(x,z_+(x),t)-\epsilon^2 z'_+(x)D_x(x,z_+(x))\bar{p}_x(x,z_+(x),t)=0$$
$$\text{for all } x\in(-1,1),\ t\in(0,\infty),$$
$$D_z(x,z_-(x))\bar{p}_z(x,z_-(x),t)-\epsilon^2 z'_-(x)D_x(x,z_-(x))\bar{p}_x(x,z_-(x),t)=0$$
$$\text{for all } x\in(-1,1),\ t\in(0,\infty),$$
$$\bar{p}(x,z,0)=\bar{p}_0(x,z) \quad \text{for all } (x,z)\in\bar{M}'.$$

The partial differential equation (2.22) is linear and strongly parabolic, with boundary conditions of nondegenerate (via (2.1)) weighted Neumann type, after which existence and uniqueness follow from the classical theory in [7, Chapter 3]. Equation (2.21) follows from an application of Green's theorem on the plane to (2.22) in $\bar{M}'$ on using the associated boundary conditions and regularity in (i)–(iii). □

The steady state problem associated with [IBVP] is

$$\hat{u}_x+\hat{w}_z=\sum_{i=1}^{N}s_i(z)\delta(x-x_i), \quad (x,z)\in M', \tag{2.23}$$
$$\hat{u}=-D_x(x,z)\hat{p}_x, \quad (x,z)\in M', \tag{2.24}$$
$$\epsilon^2\hat{w}=-D_z(x,z)\hat{p}_z, \quad (x,z)\in M', \tag{2.25}$$
$$\hat{u}(-1,z)=0 \quad \text{for all } z\in[z_-(-1),z_+(-1)], \tag{2.26}$$
$$\hat{u}(1,z)=0 \quad \text{for all } z\in[z_-(1),z_+(1)], \tag{2.27}$$
$$\hat{w}(x,z_+(x))-z'_+(x)\hat{u}(x,z_+(x))=0 \quad \text{for all } x\in(-1,1), \tag{2.28}$$
$$\hat{w}(x,z_-(x))-z'_-(x)\hat{u}(x,z_-(x))=0 \quad \text{for all } x\in(-1,1), \tag{2.29}$$

which we will refer to as [SSP]. Corresponding to (i)–(iii), a solution to [SSP] has the following regularity:

(si) $\hat{p}\in C(\bar{M}')\cap C^1(\bar{M}'\backslash\bar{d})\cap C^2(M'\backslash d)$, $\hat{u}\in C(\bar{M}'\backslash\bar{d})\cap C^1(M'\backslash d)$, $\hat{w}\in C(\bar{M}')\cap C^1(M'\backslash d)$;

(sii) $\lim_{x\to x_i^\pm}\hat{p}_x$ and $\lim_{x\to x_i^\pm}\hat{p}_z$ both exist uniformly for $z\in[z_-(x_i),z_+(x_i)]$ for each $i=1,\dots,N$;

(siii) $[\hat{p}_z]_{x_i^-}^{x_i^+}=0$, $[\hat{p}_x]_{x_i^-}^{x_i^+}=-s_i(z)/D_x(x_i,z)$ for all $z\in[z_-(x_i),z_+(x_i)]$ and each $i=1,\dots,N$.

As before, $\lim_{x\to x_i^\pm}\hat{u}$ and $\lim_{x\to x_i^\pm}\hat{w}$ exist uniformly for $z\in[z_-(x_i),z_+(x_i)]$, and

$$[\hat{u}]_{x_i^-}^{x_i^+}=s_i(z), \quad [\hat{w}]_{x_i^-}^{x_i^+}=0 \tag{2.30}$$

for all $z\in[z_-(x_i),z_+(x_i)]$ and for each $i=1,\dots,N$. Again, following standard theory for linear strongly elliptic weighted Neumann problems (see, for example, [7] or [17, Chapters 8, 9]), we have the following result.

Theorem 2.2. *For each $\epsilon > 0$,* [SSP] *has a unique (up to addition of a constant in $\hat{p}$) solution $\hat{u}, \hat{w}, \hat{p} : \bar{M}' \mapsto \mathbb{R}$ if and only if*

$$\sum_{i=1}^{N} \alpha_i = 0. \tag{2.31}$$

In what follows we will assume that the specified flux constants α_i, $i = 1, \dots, N$, satisfy condition (2.31). Now let $\tilde{u}, \tilde{w}, \tilde{p} : \bar{M}' \times [0, \infty) \mapsto \mathbb{R}$ be defined by $\tilde{u} = u - \hat{u}$, $\tilde{w} = w - \hat{w}$, $\tilde{p} = \bar{p} - \hat{p}$. It then follows that $\tilde{u}, \tilde{w}$, and $\tilde{p}$ are solutions to the problem

$$\tilde{p}_t - \left\{ \left(D_x(x,z)\tilde{p}_x\right)_x + \left(\epsilon^{-2} D_z(x,z)\tilde{p}_z\right)_z \right\} = 0,\ (x,z) \in M',\ t \in (0,\infty), \tag{2.32}$$

$$\tilde{p}_x(-1,z,t) = 0 \text{ for all } z \in [z_-(-1), z_+(-1)],\ t \in (0,\infty), \tag{2.33}$$

$$\tilde{p}_x(1,z,t) = 0 \text{ for all } z \in [z_-(1), z_+(1)],\ t \in (0,\infty), \tag{2.34}$$

$$\begin{gathered} D_z(x, z_+(x))\tilde{p}_z(x, z_+(x), t) - \epsilon^2 z_+'(x) D_x(x, z_+(x))\tilde{p}_x(x, z_+(x), t) = 0 \\ \text{for all } x \in (-1,1),\ t \in (0,\infty), \end{gathered} \tag{2.35}$$

$$\begin{gathered} D_z(x, z_-(x))\tilde{p}_z(x, z_-(x), t) - \epsilon^2 z_-'(x) D_x(x, z_-(x))\tilde{p}_x(x, z_-(x), t) = 0 \\ \text{for all } x \in (-1,1),\ t \in (0,\infty), \end{gathered} \tag{2.36}$$

$$\tilde{p}(x,z,0) = \bar{p}_0(x,z) - \hat{p}(x,z) = \tilde{p}_0(x,z) \quad \text{for all } (x,z) \in \bar{M}', \tag{2.37}$$

with regularity

$$\tilde{p} \in C(\bar{M}' \times [0,\infty)) \cap C^1(\bar{M}' \times (0,\infty)) \cap C^2(M' \times (0,\infty)), \tag{2.38}$$

after which

$$\tilde{u} = -D_x(x,z)\tilde{p}_x, \quad \tilde{w} = -\epsilon^{-2} D_z(x,z)\tilde{p}_z, \quad (x,z) \in \bar{M}' \times (0,\infty). \tag{2.39}$$

To fix the indeterminate constant in Theorem 2.2, we will take $\hat{p} : \bar{M}' \mapsto \mathbb{R}$ to be that steady state which satisfies

$$\iint_{\bar{M}'} \hat{p}(x,z)\,\mathrm{d}x\,\mathrm{d}z = \iint_{\bar{M}'} \bar{p}_0(x,z)\,\mathrm{d}x\,\mathrm{d}z =: I_0, \tag{2.40}$$

so that, via (2.37),

$$\iint_{\bar{M}'} \tilde{p}_0(x,z)\,\mathrm{d}x\,\mathrm{d}z = 0. \tag{2.41}$$

Now it follows from Theorem 2.1 that the strongly parabolic problem (2.32)–(2.39) has a unique solution in $\bar{M}' \times [0,\infty)$. We will now construct this solution. To this end we first consider the following self-adjoint eigenvalue problem in $\bar{M}'$:

$$\begin{gathered} \left(D_x(x,z)\phi_x\right)_x + \left(\epsilon^{-2} D_z(x,z)\phi_z\right)_z + \lambda\phi = 0, \quad (x,z) \in M', \\ \phi_x(-1,z) = 0 \quad \text{for all } z \in [z_-(-1), z_+(-1)], \\ \phi_x(1,z) = 0 \quad \text{for all } z \in [z_-(1), z_+(1)], \\ D_z(x, z_+(x))\phi_z(x, z_+(x)) - \epsilon^2 z_+'(x) D_x(x, z_+(x))\phi_x(x, z_+(x)) = 0 \text{ for all } x \in (-1,1), \\ D_z(x, z_-(x))\phi_z(x, z_-(x)) - \epsilon^2 z_-'(x) D_x(x, z_-(x))\phi_x(x, z_-(x)) = 0 \text{ for all } x \in (-1,1). \end{gathered}$$

We will denote this eigenvalue problem by [EVP], with $\lambda \in \mathbb{C}$ being the eigenvalue parameter. It follows from (2.1) that this is a regular, self-adjoint eigenvalue problem.

It then follows from standard theory that the eigenvalues of [EVP] are all real and given by $\lambda = \lambda_j(\epsilon)$, $j = 0, 1, 2, \ldots$, with

$$0 = \lambda_0(\epsilon) < \lambda_1(\epsilon) < \lambda_2(\epsilon) < \cdots \tag{2.42}$$

and $\lambda_j(\epsilon) \to +\infty$ as $j \to \infty$. Each corresponding eigenspace is spanned by a single eigenfunction $\phi_j : \bar{M}' \mapsto \mathbb{R}$, $j = 0, 1, 2, \ldots$, with

$$\phi_0(x, z, \epsilon) = (\text{meas}(\bar{M}'))^{-1/2} \quad \text{for all } (x, z) \in \bar{M}', \tag{2.43}$$

and meas$(\bar{M}')$ being the measure (area) of $\bar{M}' \subset \mathbb{R}^2$. Moreover,

$$\langle \phi_i, \phi_j \rangle = \iint_{\bar{M}'} \phi_i(x, z, \epsilon)\phi_j(x, z, \epsilon)\, \mathrm{d}x\, \mathrm{d}z = \delta_{ij},$$

for $i, j = 0, 1, 2, \ldots$, and δ_{ij} being the Kronecker delta symbol. Moreover, any function $\psi : \bar{M}' \mapsto \mathbb{R}$ such that $\psi \in C(\bar{M}') \cap PC^1(\bar{M}')$ and satisfies the same boundary conditions on $\partial M'$ as the set of eigenfunctions has the representation

$$\psi(x, z) = \sum_{r=0}^{\infty} \psi_r(\epsilon)\phi_r(x, z, \epsilon), \quad (x, z) \in \bar{M}', \tag{2.44}$$

with the convergence of the sum being uniform and absolute for $(x, z) \in \bar{M}'$, where

$$\psi_j(\epsilon) = \langle \psi, \phi_j \rangle = \iint_{\bar{M}'} \psi(x, z)\phi_j(x, z, \epsilon)\, \mathrm{d}x\, \mathrm{d}z, \tag{2.45}$$

for $j = 0, 1, 2, \ldots$ (see, for example, [17, Chapters 8, 9]). It is now straightforward to establish that the (unique) solution to (2.32)–(2.37) is given by

$$\tilde{p}(x, z, t) = \sum_{n=1}^{\infty} a_n(\epsilon)\mathrm{e}^{-\lambda_n(\epsilon)t}\phi_n(x, z, \epsilon), \quad (x, z) \in \bar{M}',\ t \in [0, \infty), \tag{2.46}$$

with $a_0(\epsilon) = 0$, via (2.37), (2.41), (2.43), (2.44), and (2.45), and

$$a_n(\epsilon) = \iint_{\bar{M}'} \tilde{p}_0(u, v)\phi_n(u, v, \epsilon)\, \mathrm{d}u\, \mathrm{d}v \tag{2.47}$$

for $n = 1, 2, \ldots$. We observe immediately from (2.46), with (2.42), that $\tilde{p}(x, z, t) \to 0$ as $t \to \infty$, uniformly for all $(x, z) \in \bar{M}'$, and also that $\tilde{p}_x(x, z, t), \tilde{p}_z(x, z, t) \to 0$ as $t \to \infty$, uniformly for all $(x, z) \in \bar{M}'$. Thus, we have established the following result.

THEOREM 2.3. *Let α_i, $i = 1, \ldots, N$, be such that $\sum_{i=1}^{N} \alpha_i = 0$. Then for each $\epsilon > 0$,* [IBVP] *has a unique solution $u, w, \bar{p} : \bar{M}' \times [0, \infty) \mapsto \mathbb{R}$ given by*

$$\begin{aligned}
\bar{p}(x, z, t) &= \hat{p}(x, z) + \tilde{p}(x, z, t),\\
u(x, z, t) &= \hat{u}(x, z) - D_x(x, z)\tilde{p}_x(x, z, t),\\
w(x, z, t) &= \hat{w}(x, z) - \epsilon^{-2}D_z(x, z)\tilde{p}_z(x, z, t),
\end{aligned}$$

for all $(x, z) \in \bar{M}'$ and $t \in [0, \infty)$. Here $\tilde{p} : \bar{M}' \times [0, \infty) \mapsto \mathbb{R}$ is given by (2.46), (2.47) *and $\hat{u}, \hat{w}, \hat{p} : \bar{M}' \mapsto \mathbb{R}$ is that solution to* [SSP] *which satisfies* (2.40). *Moreover,*

$$\bar{p}(x, z, t) \to \hat{p}(x, z), \quad u(x, z, t) \to \hat{u}(x, z), \quad w(x, z, t) \to \hat{w}(x, z) \quad \textit{as } t \to \infty$$

uniformly for $(x,z) \in \bar{M}'$.

We remark that since $\hat{p} \in C(\bar{M}') \cap PC^1(\bar{M}')$ and the initial data for [IBVP] $\bar{p}_0 \in C(\bar{M}') \cap PC^1(\bar{M}')$, then Theorem 2.3 implies global asymptotic stability (up to the addition of a constant to $\hat{p}$) for [SSP] with respect to perturbations in $C(\bar{M}') \cap PC^1(\bar{M}')$.

To complete the solution to the problem we are required to determine $\lambda_n(\epsilon)$ (> 0) and its corresponding eigenfunction $\phi_n : \bar{M}' \mapsto \mathbb{R}$ for each $n = 1, 2, 3, \dots$, together with the steady state $\hat{p}, \hat{u}, \hat{w} : \bar{M}' \mapsto \mathbb{R}$ which satisfies the constraint (2.40). In the next sections we focus attention on the study of [SSP] and [EVP] in turn.

In particular for a thin porous layer the parameter ϵ, which measures the aspect ratio of the layer, is small, so that $0 < \epsilon \ll 1$. In the next two sections we will consider the structure of the solutions to [SSP] and [EVP] in the asymptotic limit $\epsilon \to 0$, via the method of matched asymptotic expansions.

3. Asymptotic solution to the steady state problem [SSP] as $\boldsymbol{\epsilon \to 0}$. In this section we develop the uniform asymptotic structure of the solution to the steady state problem [SSP] (given by (2.23)–(2.29)) in the limit $\epsilon \to 0$, via the method of matched asymptotic expansions. We recall that existence and uniqueness, for each $\epsilon > 0$, follows from Theorem 2.2, and following Theorem 2.3, we require the solution of [SSP] that satisfies the constraint (2.40). Due to the initial scalings in the nondimensionalization (2.12), we anticipate that $\hat{u}, \hat{w}, \hat{p} : \bar{M}' \mapsto \mathbb{R}$ are such that

$$\hat{u}, \hat{w}, \hat{p} = O(1) \tag{3.1}$$

as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{M}' \backslash \bigcup_{i=1}^{N} \delta_i^\epsilon = \bar{N}'_\epsilon$, where δ_i^ϵ is an O(ϵ) neighborhood of $\bar{d}_i$, for each $i = 1, \dots, N$. Therefore, following (3.1), we introduce the outer region (this being $\bar{N}'_\epsilon$) asymptotic expansions

$$\begin{aligned} \hat{u}(x,z;\epsilon) &= \hat{u}_0(x,z) + \epsilon \hat{u}_1(x,z) + O(\epsilon^2), \\ \hat{w}(x,z;\epsilon) &= \hat{w}_0(x,z) + \epsilon \hat{w}_1(x,z) + O(\epsilon^2), \\ \hat{p}(x,z;\epsilon) &= \hat{p}_0(x,z) + \epsilon \hat{p}_1(x,z) + O(\epsilon^2), \end{aligned} \tag{3.2}$$

as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{N}'_\epsilon$. We substitute from (3.2) into [SSP] and condition (2.40). At leading order we obtain the following problem for $\hat{u}_0, \hat{w}_0, \hat{p}_0 : \bar{M}' \mapsto \mathbb{R}$:

$$\hat{u}_{0x} + \hat{w}_{0z} = \sum_{i=1}^{N} s_i(z) \delta(x - x_i), \quad (x,z) \in M', \tag{3.3}$$

$$\hat{u}_0 = -D_x(x,z) \hat{p}_{0x}, \quad (x,z) \in M', \tag{3.4}$$

$$0 = -D_z(x,z) \hat{p}_{0z}, \quad (x,z) \in M', \tag{3.5}$$

$$\hat{u}_0(-1,z) = 0, \quad z \in [z_-(-1), z_+(-1)], \tag{3.6}$$

$$\hat{u}_0(1,z) = 0, \quad z \in [z_-(1), z_+(1)], \tag{3.7}$$

$$\hat{w}_0(x, z_+(x)) - z'_+(x) \hat{u}_0(x, z_+(x)) = 0, \quad x \in (-1,1), \tag{3.8}$$

$$\hat{w}_0(x, z_-(x)) - z'_-(x) \hat{u}_0(x, z_-(x)) = 0, \quad x \in (-1,1), \tag{3.9}$$

$$\iint_{\bar{M}'} \hat{p}_0(x,z) \, \mathrm{d}x \, \mathrm{d}z = I_0. \tag{3.10}$$

We now construct the solution of (3.3)–(3.10). As a consequence of (2.1), equation (3.5) requires

$$\hat{p}_0(x,z) = A(x), \quad (x,z) \in \bar{M}', \tag{3.11}$$

with $A : [-1, 1] \mapsto \mathbb{R}$ to be determined. Equation (3.4) then gives

$$\hat{u}_0(x, z) = -D_x(x, z)A'(x), \quad (x, z) \in \bar{M}', \tag{3.12}$$

and the boundary conditions (3.6) and (3.7) then require $A'(-1) = A'(1) = 0$. We next substitute from (3.12) into (3.3), which becomes

$$\hat{w_0}_z = \sum_{i=1}^{N} s_i(z)\delta(x - x_i) + [D_x(x, z)A'(x)]_x, \quad (x, z) \in M'. \tag{3.13}$$

A direct integration of (3.13), together with an application of (3.9), leads to

$$\begin{aligned}\hat{w}_0(x, z) = &\sum_{i=1}^{N} F_i(z)\delta(x - x_i) \\ &+ \int_{z_-(x)}^{z} [D_x(x, \lambda)A'(x)]_x \,\mathrm{d}\lambda - z_-'(x)D_x(x, z_-(x))A'(x), \quad (x, z) \in \bar{M}',\end{aligned} \tag{3.14}$$

where

$$F_i(z) = \int_{z_-(x_i)}^{z} s_i(\lambda)\,\mathrm{d}\lambda, \quad z \in [z_-(x_i), z_+(x_i)], \tag{3.15}$$

for each $i = 1, \dots, N$. (Note that $F_i : [z_-(x_i), z_+(x_i)] \mapsto \mathbb{R}$ is such that $F_i \in C^1([z_-(x_i), z_+(x_i)])$, for each $i = 1, \dots, N$.) It remains to apply the boundary condition (3.8). The application of (3.8) using (3.12) and (3.15) finally requires that

$$\begin{aligned}&\int_{z_-(x)}^{z_+(x)} [D_x(x, \lambda)A'(x)]_x \,\mathrm{d}\lambda + \{z_+'(x)D_x(x, z_+(x)) - z_-'(x)D_x(x, z_-(x))\}A'(x) \\ &\qquad + \sum_{i=1}^{N} \alpha_i\delta(x - x_i) = 0, \quad x \in (-1, 1).\end{aligned} \tag{3.16}$$

We now rewrite the first term on the left-hand side of (3.16) as

$$\begin{aligned}&\int_{z_-(x)}^{z_+(x)} [D_x(x, \lambda)A'(x)]_x \,\mathrm{d}\lambda \\ &= \left(\int_{z_-(x)}^{z_+(x)} D_x(x, \lambda)A'(x)\,\mathrm{d}\lambda\right)' - \left\{z_+'(x)D_x(x, z_+(x)) - z_-'(x)D_x(x, z_-(x))\right\} A'(x) \\ &= (\bar{D}_x(x)A'(x))' - \left\{z_+'(x)D_x(x, z_+(x)) - z_-'(x)D_x(x, z_-(x))\right\} A'(x), \quad x \in (-1, 1).\end{aligned} \tag{3.17}$$

On substitution from (3.17) into (3.16) we obtain

$$(\bar{D}_x(x)A'(x))' = -\sum_{i=1}^{N} \alpha_i\delta(x - x_i), \quad x \in (-1, 1),$$

with

$$\bar{D}_x(x) = \int_{z_-(x)}^{z_+(x)} D_x(x, \lambda)\,\mathrm{d}\lambda, \quad x \in [-1, 1], \tag{3.18}$$

which represents the depth integrated permeability of the layer in the x-direction at each location $x \in [-1,1]$. We observe that $\bar{D}_x : [-1,1] \mapsto \mathbb{R}$ is such that $\bar{D}_x \in C^1([-1,1])$ and $\bar{D}_x(x) \geq \bar{D}_0 > 0$ for all $x \in [-1,1]$, via (2.1) and (2.2), for some positive constant $\bar{D}_0$. Thus $A : [-1,1] \mapsto \mathbb{R}$ is determined as the solution to the linear, inhomogeneous, boundary value problem (hereafter referred to as [BVP]),

$$[\bar{D}_x(x)A'(x)]' = -\sum_{i=1}^{N} \alpha_i \delta(x - x_i), \quad x \in (-1,1),$$
$$A'(-1) = A'(1) = 0,$$
$$\int_{-1}^{1} (z_+(x) - z_-(x))A(x)\,\mathrm{d}x = I_0,$$

with the final constraint arising via (3.10) on using (3.11). We observe the following.

REMARK 3.1. [BVP] *has a unique solution if and only if* $\sum_{i=1}^{N} \alpha_i = 0$.

This is in accord with condition (2.31) of Theorem 2.2. We now construct the solution to [BVP] (under condition (2.31)). It is straightforward to establish that the solution to [BVP] (with the usual Dirac delta function formalism) is given by

$$A(x) = \int_{-1}^{x} \frac{S(\lambda)}{\bar{D}_x(\lambda)}\,\mathrm{d}\lambda + A_0, \quad x \in [-1,1], \tag{3.19}$$

where here the function $S : [-1,1] \mapsto \mathbb{R}$ is the step function, given by

$$S(\lambda) = -\sum_{i=0}^{k} \alpha_i \quad \text{for all } \lambda \in [x_k, x_{k+1}), \tag{3.20}$$

and for each $k = 0, \ldots, N$, where we have defined $\alpha_0 = 0$, $x_0 = -1$, $x_{N+1} = 1$. The constant $A_0 \in \mathbb{R}$ is given by

$$A_0 = \frac{I_0}{\mathrm{meas}(\bar{M}')} - \frac{1}{\mathrm{meas}(\bar{M}')} \int_{-1}^{1} \frac{S(\lambda)\,\mathrm{meas}(\bar{M}'(\lambda))}{\bar{D}_x(\lambda)}\,\mathrm{d}\lambda, \tag{3.21}$$

where

$$\mathrm{meas}(\bar{M}'(\lambda)) = \int_{\lambda}^{1} (z_+(\mu) - z_-(\mu))\,\mathrm{d}\mu \quad \text{for all } \lambda \in [-1,1],$$

so that $\mathrm{meas}(\bar{M}'(-1)) = \mathrm{meas}(\bar{M}')$. We observe that $A \in C([-1,1]) \cap PC^2([-1,1])$, and that $A'(x_j^+) - A'(x_j^-) = -\alpha_j/\bar{D}_x(x_j)$ for each $j = 1, \ldots, N$. We can now reconstruct the solution to the leading order problem as

$$\hat{p}_0(x,z) = A(x), \quad (x,z) \in \bar{M}', \tag{3.22}$$

$$\hat{u}_0(x,z) = \frac{-D_x(x,z)}{\bar{D}_x(x)} S(x), \quad (x,z) \in \bar{M}', \tag{3.23}$$

$$\hat{w}_0(x,z) = S(x) \int_{z_-(x)}^{z} \left\{ \frac{D_x(x,\lambda)}{\bar{D}_x(x)} \right\}_x \mathrm{d}\lambda - \frac{z_-'(x) D_x(x, z_-(x)) S(x)}{\bar{D}_x(x)}, \quad (x,z) \in \bar{N}_\epsilon', \tag{3.24}$$

from (3.11), (3.12), (3.14), and (3.19). It follows from (3.22)–(3.24) that $\hat{p}_0 \in C(\bar{M}') \cap C^1(\bar{M}'\backslash\bar{d}) \cap C^2(M'\backslash d)$, $\hat{u}_0 \in C(\bar{M}'\backslash\bar{d}) \cap C^1(M'\backslash d)$, $\hat{w}_0 \in C(\bar{M}'\backslash\bar{d}) \cap C^1(M'\backslash d)$, and

$$[\hat{p}_{0z}]_{x_i^-}^{x_i^+} = 0, \tag{3.25}$$

$$[\hat{p}_{0x}]_{x_i^-}^{x_i^+} = \frac{-\alpha_i}{\bar{D}_x(x_i)}, \tag{3.26}$$

$$[\hat{u}_0]_{x_i^-}^{x_i^+} = \frac{D_x(x_i,z)\alpha_i}{\bar{D}_x(x_i)}, \tag{3.27}$$

$$[\hat{w}_0]_{x_i^-}^{x_i^+} = \left\{ \frac{z_-'(x_i)D_x(x_i,z_-(x_i))}{\bar{D}_x(x_i)} - \int_{z_-(x_i)}^{z} \left[\left\{ \frac{D_x(x,\lambda)}{\bar{D}_x(x)} \right\}_x \right]_{x=x_i} \mathrm{d}\lambda \right\} \alpha_i \tag{3.28}$$

for $z \in [z_-(x_i), z_+(x_i)]$ and for each $i = 1,\ldots,N$. We now proceed to $O(\epsilon)$. The problem for $\hat{u}_1, \hat{w}_1, \hat{p}_1 : \bar{M}' \mapsto \mathbb{R}$ is similar to the leading order problem and is not repeated here. We obtain

$$\begin{gathered} \hat{p}_1(x,z) = B(x), \quad (x,z) \in \bar{M}', \\ \hat{u}_1(x,z) = -D_x(x,z)B'(x), \quad (x,z) \in \bar{M}', \\ \hat{w}_1(x,z) = \int_{z_-(x)}^{z} [D_x(x,\lambda)B'(x)]_x \,\mathrm{d}\lambda - z_-'(x)D_x(x,z_-(x))B'(x), \quad (x,z) \in \bar{M}', \end{gathered} \tag{3.29}$$

where $B : [-1,1] \mapsto \mathbb{R}$ is the solution to the boundary value problem

$$\begin{gathered} [\bar{D}_x(x)B'(x)]' = 0, \quad x \in (-1,1), \\ B'(-1) = B'(1) = 0, \\ \int_{-1}^{1} (z_+(x) - z_-(x))B(x)\,\mathrm{d}x = 0. \end{gathered}$$

The unique solution $B \in C^1([-1,1]) \cap C^2((-1,1))$ is given by $B(x) = 0$ for all $x \in [-1,1]$, and so $\hat{p}_1(x,z) = \hat{u}_1(x,z) = \hat{w}_1(x,z) = 0$ for $(x,z) \in \bar{M}'$, via (3.29). The outer region asymptotic expansions are thus

$$\hat{u}(x,z;\epsilon) = \frac{-D_x(x,z)}{\bar{D}_x(x)}S(x) + O(\epsilon^2), \tag{3.30}$$

$$\hat{w}(x,z;\epsilon) = S(x)\int_{z_-(x)}^{z} \left\{ \frac{D_x(x,\lambda)}{\bar{D}_x(x)} \right\}_x \mathrm{d}\lambda - \frac{z_-'(x)D_x(x,z_-(x))S(x)}{\bar{D}_x(x)} + O(\epsilon^2), \tag{3.31}$$

$$\hat{p}(x,z;\epsilon) = A(x) + O(\epsilon^2), \tag{3.32}$$

as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{N}_\epsilon'$, with $A, S : [-1,1] \mapsto \mathbb{R}$ given by (3.19)–(3.21).

We now observe from (3.30)–(3.32), via (3.25)–(3.28), that all of the regularity requirements in (si), together with the limit conditions (sii), (siii), and (2.30), are not satisfied at $x = x_i$ for each $z \in [z_-(x_i), z_+(x_i)]$, with $i = 1,\ldots,N$ (although the integrated forms are satisfied). We conclude (as was anticipated earlier) that the outer region asymptotic expansions (3.30)–(3.32) become nonuniform when $(x,z) \in \delta_i^\epsilon$ as $\epsilon \to 0$, $i = 1,\ldots,N$. To obtain a uniform asymptotic representation to the solution to [SSP] when $(x,z) \in \delta_i^\epsilon$ as $\epsilon \to 0$, we must therefore introduce an inner region at each line source/sink location $x = x_i$, $i = 1,\ldots,N$. We now consider the inner region in the neighborhood of $x = x_i$ in detail. In the inner region, $x = x_i + O(\epsilon)$, $z = O(1)$, as $\epsilon \to 0$, with, from (3.30)–(3.32), $\hat{u} = O(1)$, $\hat{w} = O(\epsilon^{-1})$, $\hat{p} = A_i + O(\epsilon)$, as $\epsilon \to 0$, with $A_i = A(x_i)$, $i = 1, \dots, N$. Thus, in the inner region we write

$$x = x_i + \epsilon X, \tag{3.33}$$

with $X \in (-\infty, \infty)$ such that $X = O(1)$ as $\epsilon \to 0$, together with

$$\hat{u} = U, \quad \hat{w} = \epsilon^{-1} W, \quad \hat{p} = A_i + \epsilon P, \tag{3.34}$$

where $U, W, P : (-\infty, \infty) \times [z_-(x_i), z_+(x_i)] \mapsto \mathbb{R}$ are such that $U, W, P = O(1)$ as $\epsilon \to 0$. We now substitute from (3.33), (3.34) into the full problem [SSP] ((2.23), (2.24), (2.25), (2.28), (2.29), excluding conditions (2.26), (2.27) which lie outside the inner region in the limit $\epsilon \to 0$). The full problem in the inner region then becomes

$$U_X + W_z = s_i(z)\delta(X), \quad (X, z) \in D(\epsilon), \tag{3.35}$$

$$U = -D_x(x_i + \epsilon X, z) P_X, \quad (X, z) \in D(\epsilon), \tag{3.36}$$

$$W = -D_z(x_i + \epsilon X, z) P_z, \quad (X, z) \in D(\epsilon), \tag{3.37}$$

$$W - \epsilon z_+'(x_i + \epsilon X) U = 0, \quad X \in (-\infty, \infty), \quad z = z_+(x_i + \epsilon X), \tag{3.38}$$

$$W - \epsilon z_-'(x_i + \epsilon X) U = 0, \quad X \in (-\infty, \infty), \quad z = z_-(x_i + \epsilon X), \tag{3.39}$$

together with matching conditions to the outer region as $|X| \to \infty$. Here

$$D(\epsilon) = \{(X, z) \in \mathbb{R}^2 : X \in (-\infty, \infty) \text{ and } z \in (z_-(x_i + \epsilon X), z_+(x_i + \epsilon X))\}. \tag{3.40}$$

We now introduce the inner region asymptotic expansions as

$$\begin{aligned} U(X, z; \epsilon) &= U_0(X, z) + O(\epsilon), \\ W(X, z; \epsilon) &= W_0(X, z) + O(\epsilon), \\ P(X, z; \epsilon) &= P_0(X, z) + O(\epsilon), \end{aligned} \tag{3.41}$$

as $\epsilon \to 0$, with $(X, z) \in \bar{D}(\epsilon)$. On substitution from (3.41) into (3.35)–(3.40) we obtain the leading order problem as

$$U_{0X} + W_{0z} = s_i(z)\delta(X), \quad (X, z) \in D(0), \tag{3.42}$$

$$U_0 = -\tilde{D}_x(z) P_{0X}, \quad (X, z) \in D(0), \tag{3.43}$$

$$W_0 = -\tilde{D}_z(z) P_{0z}, \quad (X, z) \in D(0), \tag{3.44}$$

$$W_0(X, z_+^i) = 0, \quad X \in (-\infty, \infty), \tag{3.45}$$

$$W_0(X, z_-^i) = 0, \quad X \in (-\infty, \infty). \tag{3.46}$$

Here,

$$\tilde{D}_x(z) = D_x(x_i, z), \quad \tilde{D}_z(z) = D_z(x_i, z), \tag{3.47}$$

for all $z \in [z_-^i, z_+^i]$, with $z_-^i = z_-(x_i)$ and $z_+^i = z_+(x_i)$. Also, $\bar{D}(0)$, via (3.40), is now the unbounded region in the (X, z) plane contained between the coordinate lines $z = z_-^i$ and $z = z_+^i$; that is,

$$\bar{D}(0) = (-\infty, \infty) \times [z_-^i, z_+^i]. \tag{3.48}$$

The leading order problem (3.42)–(3.46) is completed by applying the asymptotic matching principle of Van Dyke [20]. It is straightforward to establish that matching of $\hat{p}$ is sufficient, after which matching of $\hat{u}$ and $\hat{w}$ follows automatically. We must apply Van Dyke's matching principle to the outer region asymptotic expansion for $\hat{p}$ taken to $O(\epsilon)$, (3.32), with the inner region asymptotic expansion for $\hat{p}$ taken to $O(\epsilon)$, (3.34) and (3.41). The appropriate matching condition is readily determined as

$$P_0(X,z) = A_i^{\pm\prime} X + o(1) \quad \text{as } X \to \pm\infty, \text{ uniformly for } z \in [z_-^i, z_+^i], \tag{3.49}$$

with

$$A_i^{+\prime} = A'(x_i^+) = -\frac{\sum_{j=0}^{i} \alpha_j}{\bar{D}_x(x_i)},\ A_i^{-\prime} = A'(x_i^-) = -\frac{\sum_{j=0}^{i-1} \alpha_j}{\bar{D}_x(x_i)},\ \text{for } i = 1, \dots, N, \tag{3.50}$$

via (3.19) and (3.20). Finally the regularity conditions (si)–(siii) with (2.30) require the following:

(Ii) $P_0 \in C(\bar{D}(0)) \cap C^1(\bar{D}(0)\backslash\bar{I}) \cap C^2(D(0)\backslash I)$, $U_0 \in C(\bar{D}(0)\backslash\bar{I}) \cap C^1(D(0)\backslash I)$, $W_0 \in C(\bar{D}(0)) \cap C^1(D(0)\backslash I)$, where $I = \{0\} \times (z_-^i, z_+^i)$;

(Iii) $\lim_{X\to 0^\pm} P_{0X}$ and $\lim_{X\to 0^\pm} P_{0z}$ both exist uniformly for $z \in [z_-^i, z_+^i]$;

(Iiii) $[P_{0z}]_{0^-}^{0^+} = 0$, $[P_{0X}]_{0^-}^{0^+} = -s_i(z)/\tilde{D}_x(z)$, $[U_0]_{0^-}^{0^+} = s_i(z)$, and $[W_0]_{0^-}^{0^+} = 0$ for all $z \in [z_-^i, z_+^i]$.

We can now eliminate U_0 and W_0, via (3.43) and (3.44), and obtain the following strongly elliptic problem for P_0, namely,

$$(\tilde{D}_x(z)P_{0X})_X + (\tilde{D}_z(z)P_{0z})_z = -s_i(z)\delta(X), \quad (X,z) \in D(0), \tag{3.51}$$

$$P_{0z}(X, z_+^i) = 0, \quad X \in (-\infty,\infty), \tag{3.52}$$

$$P_{0z}(X, z_-^i) = 0, \quad X \in (-\infty,\infty), \tag{3.53}$$

$$P_0(X,z) = A_i^{\pm\prime} X + o(1), \quad X \to \pm\infty, \quad \text{uniformly for } z \in [z_-^i, z_+^i], \tag{3.54}$$

together with (Ii)–(Iiii). The first step in obtaining the solution to (3.51)–(3.54) is to consider the regular Sturm–Liouville eigenvalue problem,

$$\begin{gathered}(\tilde{D}_z(z)\psi_z)_z + \bar{\lambda}\tilde{D}_x(z)\psi = 0, \quad z \in (z_-^i, z_+^i),\\ \psi_z(z_-^i) = \psi_z(z_+^i) = 0,\end{gathered}$$

which we refer to as [SL]. Here $\bar{\lambda} \in \mathbb{C}$ is the eigenvalue parameter. Classical Sturm–Liouville theory (see, for example, [6, Chapters 7, 8]) determines that the set of eigenvalues of [SL] is given by $\bar{\lambda} = \bar{\lambda}_r \in \mathbb{R}$, $r = 0, 1, 2, \dots$, with $0 = \bar{\lambda}_0 < \bar{\lambda}_1 < \bar{\lambda}_2 < \bar{\lambda}_3 < \cdots$, where $\bar{\lambda}_r \to +\infty$ as $r \to \infty$. The corresponding normalized eigenfunctions $\psi_r : [z_-^i, z_+^i] \mapsto \mathbb{R}$ form an orthonormal set, so that

$$\langle \psi_r, \psi_s \rangle = \int_{z_-^i}^{z_+^i} \tilde{D}_x(z)\psi_r(z)\psi_s(z)\,\mathrm{d}z = \delta_{rs} \quad \text{for } r, s = 0, 1, 2, \dots. \tag{3.55}$$

The set of eigenfunctions of [SL] are complete on the interval $[z_-^i, z_+^i]$. Completeness allows us to write the solution to (3.51) with conditions (3.52), (3.53) as

$$P_0(X,z) = \begin{cases} \sum_{n=0}^{\infty} \chi_n^+(X)\psi_n(z), & X > 0,\\ \sum_{n=0}^{\infty} \chi_n^-(X)\psi_n(z), & X < 0,\end{cases} \tag{3.56}$$

with $z \in [z_-^i, z_+^i]$. Substitution of (3.56) into (3.51) establishes that

$$\begin{aligned}\chi_n^+(X) &= A_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} + B_n \mathrm{e}^{-\bar{\lambda}_n^{1/2} X}, \quad X > 0,\\ \chi_n^-(X) &= D_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} + C_n \mathrm{e}^{-\bar{\lambda}_n^{1/2} X}, \quad X < 0,\end{aligned} \tag{3.57}$$

with $A_n, B_n, C_n, D_n \in \mathbb{R}$ constants, for $n = 1, 2, \ldots$. With $n = 0$, we have

$$\begin{aligned} \chi_0^+(X) &= A_0 + B_0 X, \quad X > 0, \\ \chi_0^-(X) &= C_0 + D_0 X, \quad X < 0, \end{aligned} \tag{3.58}$$

with $A_0, B_0, C_0, D_0 \in \mathbb{R}$ constants. At this stage we observe that

$$\psi_0(z) = \left\{ \int_{z_-^i}^{z_+^i} \tilde{D}_x(s)\,\mathrm{d}s \right\}^{-1/2} = \Psi_0, \quad z \in [z_-^i, z_+^i], \tag{3.59}$$

with $\Psi_0 > 0$. It then follows, via (3.56)–(3.59), that the boundary conditions (3.49) are satisfied if and only if

$$\begin{aligned} A_n &= 0, \quad n = 0, 1, 2, \ldots, \quad B_0 = A_i^{+\prime} \Psi_0^{-1}, \\ C_n &= 0, \quad n = 0, 1, 2, \ldots, \quad D_0 = A_i^{-\prime} \Psi_0^{-1}. \end{aligned} \tag{3.60}$$

Next, across $X = 0$, continuity of P_0, together with the condition $[P_{0z}]_{0^-}^{0^+} = 0$ of (Iiii), is satisfied if and only if, via (3.56)–(3.60),

$$B_n = D_n, \quad n = 1, 2, \ldots. \tag{3.61}$$

Finally, it remains to satisfy the condition $[P_{0X}]_{0^-}^{0^+} = -s_i(z)/\tilde{D}_x(z)$ of (Iiii), which requires, via (3.56)–(3.61), that

$$(A_i^{+\prime} - A_i^{-\prime}) - \sum_{n=1}^{\infty} 2\bar{\lambda}_n^{1/2} B_n \psi_n(z) = \frac{-s_i(z)}{\tilde{D}_x(z)}, \quad z \in [z_-^i, z_+^i]. \tag{3.62}$$

The completeness of the eigenfunctions $\psi_n(z)$ $(n = 0, 1, 2, \ldots)$ on the interval $[z_-^i, z_+^i]$ allows (3.62) to be satisfied uniquely, with, using (3.55),

$$A_i^{+\prime} - A_i^{-\prime} = -\frac{\int_{z_-^i}^{z_+^i} s_i(s)\,\mathrm{d}s}{\int_{z_-^i}^{z_+^i} \tilde{D}_x(s)\,\mathrm{d}s} = -\frac{\alpha_i}{\bar{D}_x(x_i)}$$

via (3.47) and (3.18), and which is automatically satisfied using (3.50), and

$$B_k = \frac{1}{2\bar{\lambda}_k^{1/2}} \int_{z_-^i}^{z_+^i} s_i(s)\psi_k(s)\,\mathrm{d}s, \quad k = 1, 2, \ldots. \tag{3.63}$$

Thus, the solution to (3.51)–(3.54), with regularity (Ii)–(Iiii), is given by

$$P_0(X, z) = \begin{cases} A_i^{+\prime} X + \sum_{n=1}^{\infty} B_n \mathrm{e}^{-\bar{\lambda}_n^{1/2} X} \psi_n(z), & X > 0, \\ A_i^{-\prime} X + \sum_{n=1}^{\infty} B_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} \psi_n(z), & X < 0, \end{cases} \tag{3.64}$$

with $z \in [z_-^i, z_+^i]$, and the coefficients B_n, $n = 1, 2, \ldots$, given by (3.63). $U_0(X, z)$ and $W_0(X, z)$ are now obtained directly from (3.43) and (3.44) as

$$U_0(X, z) = \begin{cases} -\tilde{D}_x(z) \left\{ A_i^{+\prime} - \sum_{n=1}^{\infty} \bar{\lambda}_n^{1/2} B_n \mathrm{e}^{-\bar{\lambda}_n^{1/2} X} \psi_n(z) \right\}, & X > 0, \\ -\tilde{D}_x(z) \left\{ A_i^{-\prime} + \sum_{n=1}^{\infty} \bar{\lambda}_n^{1/2} B_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} \psi_n(z) \right\}, & X < 0, \end{cases} \tag{3.65}$$

$$W_0(X, z) = \begin{cases} -\tilde{D}_z(z) \sum_{n=1}^{\infty} B_n \mathrm{e}^{-\bar{\lambda}_n^{1/2} X} \psi_n'(z), & X > 0, \\ -\tilde{D}_z(z) \sum_{n=1}^{\infty} B_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} \psi_n'(z), & X < 0, \end{cases} \tag{3.66}$$

with $z \in [z_-^i, z_+^i]$. The only remaining question is how to actually compute the eigenvalues and corresponding eigenfunctions of [SL]. If $\tilde{D}_x, \tilde{D}_z$, are constant with respect to z, then analytical solution of [SL] is trivial. More generally, since [SL] is a regular Sturm–Liouville problem, numerical methods are straightforward and very efficient and will not be discussed further here. The solution of the leading order problem is now complete.

It is of interest to obtain the expression for the pressure at the location of the ith line source/sink. From (3.34) and (3.41), this is given by $\hat{p}(x_i, z) = A_i + \epsilon P_0(0, z) + O(\epsilon^2)$, for $z \in [z_-^i, z_+^i]$. On using (3.64), this becomes

$$\hat{p}(x_i, z) = A_i + \epsilon \left(\sum_{n=1}^{\infty} B_n \psi_n(z) \right) + O(\epsilon^2) \quad \text{for } z \in [z_-^i, z_+^i]. \tag{3.67}$$

The pressure difference between the ith and jth line source/sinks is then

$$\begin{aligned} \Delta\hat{p}_{ij}(z) &= \hat{p}(x_i, z) - \hat{p}(x_j, z) \\ &= (A_i - A_j) + \epsilon \left(\sum_{n=1}^{\infty} [B_n^i \psi_n^i(z) - B_n^j \psi_n^j(z)] \right) + O(\epsilon^2) \end{aligned} \tag{3.68}$$

for $z \in [z_-^i, z_+^i]$, with superscripts i and j distinguishing evaluation at the ith and jth line source/sink, respectively. In (3.68), we recall that, via (3.19),

$$A_i - A_j = A(x_i) - A(x_j) = \int_{x_j}^{x_i} \frac{S(\lambda)}{\bar{D}_x(\lambda)}\, \mathrm{d}\lambda.$$

The asymptotic structure to the solution of [SSP] as $\epsilon \to 0$ is now complete. Two minor extensions are worthy of consideration at this stage and are given in the subsections that follow.

3.1. A line source/sink close to the boundary. In the above, the locations of the line source/sinks $x_i \in (-1, 1)$, $i = 1, \ldots, N$, are such that $(x_1 + 1)$, $(1 - x_N)$, and $x_{i+1} - x_i$ $(i = 1, \ldots, N-1)$ remain positive and finite $(O(1))$ as $\epsilon \to 0$. In this extension we consider the situation when $x_1 + 1 = O(\epsilon)$ as $\epsilon \to 0$, so that the line source/sink at $x = x_1$ lies within $O(\epsilon)$ of the layer boundary at $x = -1$. The structure of the outer region to [SSP] is unchanged. However, the inner region to [SSP] at $x = x_1$ now encompasses the boundary at $x = -1$, and so the leading order problem in this inner region, when $i = 1$, is modified. To formalize this we write

$$x_1 = -1 + \epsilon\bar{\sigma}, \tag{3.69}$$

with the constant $\bar{\sigma} > 0$. In terms of the inner coordinate X,

$$x = x_1 + \epsilon X, \tag{3.70}$$

via (3.33). Thus, via (3.69) and (3.70), in the inner region, the line source/sink is located at $X = 0$, while the layer boundary is located at $X = -\bar{\sigma}$. Without repeating details, the leading order problem in the inner region is now

$$U_{0X} + W_{0z} = s_1(z)\delta(X), \quad (X, z) \in D(0), \tag{3.71}$$

$$U_0 = -\tilde{D}_x(z) P_{0X}, \quad (X, z) \in D(0), \tag{3.72}$$

$$W_0 = -\tilde{D}_z(z) P_{0z}, \quad (X, z) \in D(0), \tag{3.73}$$

$$W_0(X, z_+^1) = 0, \quad X \in (-\bar{\sigma}, \infty), \tag{3.74}$$

$$W_0(X, z_-^1) = 0, \quad X \in (-\bar{\sigma}, \infty), \tag{3.75}$$

where now $\bar{D}(0)$ is the unbounded region in the (X,z) plane contained inside the coordinate lines $z = z_+^1$, $z = z_-^1$, and $X = -\bar{\sigma}$, so that

$$\bar{D}(0) = [-\bar{\sigma}, \infty) \times [z_-^1, z_+^1]. \tag{3.76}$$

The leading order problem (3.71)–(3.75) is completed with the condition

$$U_0(-\bar{\sigma}, z) = 0, \quad z \in [z_-^1, z_+^1], \tag{3.77}$$

$$P_0(X,z) = {A_1^{+}}'X + o(1), \text{ as } X \to +\infty, \text{ uniformly for } z \in [z_-^1, z_+^1], \tag{3.78}$$

with the latter being the matching condition to the outer region. The solution to (3.71)–(3.78) can be constructed as before, to obtain

$$P_0(X,z) = \begin{cases} {A_1^{+}}'X + \sum_{n=1}^{\infty} \bar{B}_n \mathrm{e}^{-\bar{\lambda}_n^{1/2}X} \psi_n(z), & X > 0, \\ \sum_{n=1}^{\infty} \frac{\bar{B}_n}{\cosh[\bar{\lambda}_n^{1/2}(X+\bar{\sigma})]} \cosh \bar{\lambda}_n^{1/2}(X+\bar{\sigma})\psi_n(z), & -\bar{\sigma} \le X < 0, \end{cases}$$

for $z \in [z_-^1, z_+^1]$, and with

$$\bar{B}_n = \frac{1}{\bar{\lambda}_n^{1/2}(1 + \tanh(\bar{\lambda}_n^{1/2}\bar{\sigma}))} \int_{z_-^1}^{z_+^1} s_1(s)\psi_n(s)\,\mathrm{d}s, \quad n = 1, 2, \ldots. \tag{3.79}$$

Equations (3.72)–(3.73) then give the corresponding expressions for $U_0(X,z)$ and $W_0(X,z)$. The pressure at the location of this first line source/sink is then given by

$$\hat{p}(x_1, z) = A_1 + \epsilon \left(\sum_{n=1}^{\infty} \bar{B}_n \psi_n(z) \right) + O(\epsilon^2) \tag{3.80}$$

for $z \in [z_-^1, z_+^1]$, with $\bar{B}_n$, $n = 1, 2, \ldots$, as given in (3.79). The difference in the expression for pressure at the wall close line source/sink (3.80), and at the interior line source/sink (3.67), occurs in the expressions for the sequence of constants B_n, (3.63), and $\bar{B}_n$, (3.79), $n = 1, 2, \ldots$.

3.2. Two closely located line source/sinks. In this extension, we consider the situation when the kth and $(k+1)$th line source/sinks are within $O(\epsilon)$ separation of each other. With $x_k, x_{k+1} \in (-1, 1)$, we write $x_{k+1} = x_k + \tilde{\sigma}\epsilon$, with the constant $\tilde{\sigma} > 0$. In terms of the inner coordinate X, $x = x_k + \epsilon X$, via (3.33). Thus both line source/sinks at $x = x_k$ and $x = x_{k+1}$ are located in the inner region at $x = x_k$, with their respective locations in this inner region being at $X = 0$ and $X = \tilde{\sigma}$. Without repeating details, the leading order problem in the inner region is now

$$U_{0X} + W_{0z} = s_k(z)\delta(X) + s_{k+1}(z)\delta(X - \tilde{\sigma}), \quad (X,z) \in D(0), \tag{3.81}$$

$$U_0 = -\tilde{D}_x(z)P_{0X}, \quad (X,z) \in D(0), \tag{3.82}$$

$$W_0 = -\tilde{D}_z(z)P_{0z}, \quad (X,z) \in D(0), \tag{3.83}$$

$$W_0(X, z_+^k) = 0, \quad X \in (-\infty, \infty), \tag{3.84}$$

$$W_0(X, z_-^k) = 0, \quad X \in (-\infty, \infty), \tag{3.85}$$

$$P_0(X,z) = \begin{cases} {A_{k+1}^{+}}'(X - \tilde{\sigma}) + \tilde{\sigma}{A_{k+1}^{-}}' + o(1), & X \to +\infty, \\ {A_k^{-}}'X + o(1), & X \to -\infty, \end{cases} \tag{3.86}$$
$$\text{uniformly for } z \in [z_-^k, z_+^k],$$

with $\bar{D}(0)$ as in (3.48) when $i = k$, and condition (3.86) being the appropriate matching condition to the outer region. The solution to (3.81)–(3.86) can be constructed as before, to obtain

$$P_0(X,z) = \begin{cases} {A_k^-}'X + \sum_{n=1}^{\infty} \hat{B}_n \mathrm{e}^{\bar{\lambda}_n^{1/2} X} \psi_n(z), & X < 0, \\ {A_{k+1}^-}'X + \sum_{n=1}^{\infty} (\hat{B}_n \cosh(\bar{\lambda}_n^{1/2} X) + \hat{D}_n \sinh(\bar{\lambda}_n^{1/2} X))\psi_n(z), & 0 < X < \tilde{\sigma}, \\ \{\tilde{\sigma}{A_{k+1}^-}' + {A_{k+1}^+}'(X - \tilde{\sigma})\} + \sum_{n=1}^{\infty} \hat{C}_n \mathrm{e}^{-\bar{\lambda}_n^{1/2}(X-\tilde{\sigma})} \psi_n(z), & X > \tilde{\sigma}, \end{cases} \tag{3.87}$$

with $z \in [z_-^k, z_+^k]$, and the coefficients

$$\hat{B}_n = \frac{1}{2\bar{\lambda}_n^{1/2}} \left\{ \frac{1}{[\cosh(\bar{\lambda}_n^{1/2}\tilde{\sigma}) + \sinh(\bar{\lambda}_n^{1/2}\tilde{\sigma})]} \int_{z_-^k}^{z_+^k} s_{k+1}(s)\psi_n(s)\,\mathrm{d}s + \int_{z_-^k}^{z_+^k} s_k(s)\psi_n(s)\,\mathrm{d}s \right\}, \tag{3.88}$$

$$\hat{D}_n = \frac{1}{2\bar{\lambda}_n^{1/2}} \left\{ \frac{1}{[\cosh(\bar{\lambda}_n^{1/2}\tilde{\sigma}) + \sinh(\bar{\lambda}_n^{1/2}\tilde{\sigma})]} \int_{z_-^k}^{z_+^k} s_{k+1}(s)\psi_n(s)\,\mathrm{d}s - \int_{z_-^k}^{z_+^k} s_k(s)\psi_n(s)\,\mathrm{d}s \right\},$$

$$\hat{C}_n = \frac{1}{2\bar{\lambda}_n^{1/2}} \left\{ \int_{z_-^k}^{z_+^k} s_{k+1}(s)\psi_n(s)\,\mathrm{d}s + [\cosh(\bar{\lambda}_n^{1/2}\tilde{\sigma}) - \sinh(\bar{\lambda}_n^{1/2}\tilde{\sigma})] \int_{z_-^k}^{z_+^k} s_k(s)\psi_n(s)\,\mathrm{d}s \right\}. \tag{3.89}$$

Equations (3.82)–(3.83) then give the corresponding expressions for $W_0(X,z)$ and $U_0(X,z)$. (Note that $\tilde{D}_x(z)$ and $\tilde{D}_z(z)$ in (3.82), (3.83) are evaluated at $x = x_k$.) The pressures at the line source/sinks at $x = x_k$ and $x = x_{k+1}$ are given by, via (3.87),

$$\hat{p}(x_k, z) = A_k + \epsilon \left(\sum_{n=1}^{\infty} \hat{B}_n \psi_n(z) \right) + O(\epsilon^2),$$

$$\hat{p}(x_{k+1}, z) = A_k + \epsilon \left(\tilde{\sigma}{A_{k+1}^-}' + \sum_{n=1}^{\infty} \hat{C}_n \psi_n(z) \right) + O(\epsilon^2),$$

with $\hat{B}_n$ and $\hat{C}_n$, $n = 1, 2, \ldots$, as given in (3.88) and (3.89). (Note in the above that $A_k + \epsilon\tilde{\sigma}{A_{k+1}^-}' = A_k + \epsilon\tilde{\sigma}A_k^{+'} + O(\epsilon^2) = A_{k+1} + O(\epsilon^2)$.)

The asymptotic solution to [SSP] as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{M}'$, is now complete. We now turn our attention to the eigenvalue problem [EVP].

4. Asymptotic solution to the eigenvalue problem [EVP] as $\epsilon \to 0$. In this section we develop the asymptotic solution to the eigenvalue problem [EVP] as $\epsilon \to 0$. We first employ the theory developed by Ramm [16] to establish that the set of eigenvalues to [EVP], (2.42), with $\epsilon > 0$, splits into two disjoint subsets as $\epsilon \to 0^+$, denoted by $S_- = \{\lambda_0^-(\epsilon), \lambda_1^-(\epsilon), \lambda_2^-(\epsilon), \ldots\}$ and $S_+ = \{\lambda_1^+(\epsilon), \lambda_2^+(\epsilon), \lambda_3^+(\epsilon), \ldots\}$, with $0 = \lambda_0^-(\epsilon) < \lambda_1^-(\epsilon) < \cdots$ and $0 < \lambda_1^+(\epsilon) < \lambda_2^+(\epsilon) < \cdots$. In particular,

$$\lambda_n^-(\epsilon) = O(n^2), \quad \lambda_n^+(\epsilon) = O(n^2\epsilon^{-2}) \tag{4.1}$$

as $\epsilon \to 0^+$, uniformly for $n = 1, 2, \ldots$. We will focus attention on the eigenvalues and corresponding eigenfunctions in the set S_-, so that in [EVP] we have $\lambda(\epsilon) = O(1)$ as $\epsilon \to 0^+$, via (4.1). Thus we expand $\phi : \bar{M}' \mapsto \mathbb{R}$ in the form

$$\phi(x,z;\epsilon) = \tilde{\phi}(x,z) + \epsilon^2\hat{\phi}(x,z) + o(\epsilon^2) \quad \text{as } \epsilon \to 0^+, \tag{4.2}$$

uniformly for $(x,z)\in\bar{M}'$, while we expand

$$\lambda(\epsilon)=\tilde{\lambda}+\epsilon^2\hat{\lambda}+o(\epsilon^2)\quad \text{as } \epsilon\to 0^+. \tag{4.3}$$

On substitution from (4.2) and (4.3) into [EVP], we obtain the leading order problem

$$\Big(D_z(x,z)\tilde{\phi}_z\Big)_z=0,\quad (x,z)\in M', \tag{4.4}$$

$$\tilde{\phi}_x(-1,z)=0,\quad z\in[z_-(-1),z_+(-1)], \tag{4.5}$$

$$\tilde{\phi}_x(1,z)=0,\quad z\in[z_-(1),z_+(1)], \tag{4.6}$$

$$\tilde{\phi}_z(x,z_+(x))=0,\quad x\in(-1,1), \tag{4.7}$$

$$\tilde{\phi}_z(x,z_-(x))=0,\quad x\in(-1,1). \tag{4.8}$$

A direct integration of (4.4) gives

$$\tilde{\phi}_z(x,z)=\frac{\tilde{B}(x)}{D_z(x,z)},\quad (x,z)\in\bar{M}', \tag{4.9}$$

while (4.7) and (4.8) require $\tilde{B}(x)=0$ for all $x\in[-1,1]$. Hence, from (4.9),

$$\tilde{\phi}(x,z)=\tilde{A}(x),\quad (x,z)\in\bar{M}', \tag{4.10}$$

with $\tilde{A}:[-1,1]\mapsto\mathbb{R}$ such that $\tilde{A}\in C^1([-1,1])\cap C^2((-1,1))$. Conditions (4.5) and (4.6) then require $\tilde{A}'(-1)=\tilde{A}'(1)=0$. At $O(\epsilon^2)$ we obtain the problem

$$\Big(D_z(x,z)\hat{\phi}_z\Big)_z=-\hat{\lambda}\tilde{A}(x)-\Big(D_x(x,z)\tilde{A}'(x)\Big)_x,\quad (x,z)\in M', \tag{4.11}$$

$$\hat{\phi}_x(-1,z)=0,\quad z\in[z_-(-1),z_+(-1)], \tag{4.12}$$

$$\hat{\phi}_x(1,z)=0,\quad z\in[z_-(1),z_+(1)], \tag{4.13}$$

$$D_z(x,z_+(x))\hat{\phi}_z(x,z_+(x))=z_+'(x)D_x(x,z_+(x))\tilde{A}'(x),\quad x\in(-1,1), \tag{4.14}$$

$$D_z(x,z_-(x))\hat{\phi}_z(x,z_-(x))=z_-'(x)D_x(x,z_-(x))\tilde{A}'(x),\quad x\in(-1,1). \tag{4.15}$$

The solvability requirement on the inhomogeneous boundary value problem (4.11)–(4.15) will provide the ordinary differential equation which must be satisfied by $\tilde{A}(x)$, $x\in(-1,1)$. On integrating (4.11) with respect to z (with $x\in(-1,1)$ fixed) between $z=z_-(x)$ and $z=z_+(x)$, we obtain

$$\begin{aligned}D_z(x,z_+(x))\hat{\phi}_z(x,z_+(x))-D_z(x,z_-(x))\hat{\phi}_z(x,z_-(x))&=-\tilde{\lambda}\tilde{A}(x)(z_+(x)-z_-(x))\\ -(\bar{D}_x(x)\tilde{A}'(x))'+[z_+'(x)D_x(x,z_+(x))&-z_-'(x)D_x(x,z_-(x))]\tilde{A}'(x)\end{aligned} \tag{4.16}$$

for all $x\in(-1,1)$. We next substitute into the left-hand side of (4.16) from (4.14) and (4.15), which, after cancellation, results in the ordinary differential equation

$$(\bar{D}_x(x)\tilde{A}'(x))'+\tilde{\lambda}(z_+(x)-z_-(x))\tilde{A}(x)=0,\quad x\in(-1,1). \tag{4.17}$$

Thus $\tilde{A}:[-1,1]\mapsto\mathbb{R}$ and $\tilde{\lambda}\in\mathbb{R}$ satisfy the regular Sturm–Liouville eigenvalue problem (which we denote hereafter by [SLP]),

$$\begin{gathered}(\bar{D}_x(x)\tilde{A}'(x))'+\tilde{\lambda}h(x)\tilde{A}(x)=0,\quad x\in(-1,1),\\ \tilde{A}'(-1)=\tilde{A}'(1)=0,\end{gathered}$$

with $\bar{D}_x(x)$ as defined in (3.18) and $h(x) = z_+(x) - z_-(x)$, $x \in [-1,1]$. Now, the classical Sturm–Liouville theory (see, for example, [6, Chapters 7, 8]) determines that the set of eigenvalues of [SLP] is given by $\tilde{\lambda} = \tilde{\lambda}_r$, $r = 0, 1, 2, \ldots$, with

$$0 = \tilde{\lambda}_0 < \tilde{\lambda}_1 < \tilde{\lambda}_2 < \cdots \quad \text{and} \quad \tilde{\lambda}_r = O(r^2) \quad \text{as } r \to \infty. \tag{4.18}$$

We remark also that [SLP] is identical in structure to the eigenvalue problem [SL] considered in section 3. Corresponding to each eigenvalue $\tilde{\lambda}_r$ ($r = 0, 1, 2, \ldots$), there is a unique normalized eigenfunction $\tilde{A}_r : [-1,1] \mapsto \mathbb{R}$ such that

$$\int_{-1}^{1} h(x)\tilde{A}_i(x)\tilde{A}_j(x)\,\mathrm{d}x = \delta_{ij} \quad \text{for } i,j = 0,1,2,\ldots. \tag{4.19}$$

We note that $\tilde{A}_0(x) = \{\int_{-1}^{1} h(s)\,\mathrm{d}s\}^{-1/2} = (\mathrm{meas}(\bar{M}'))^{-1/2}$ for all $x \in [-1,1]$. Thus, we have established for [EVP], via (4.2), (4.3), (4.10), that $\lambda_r^-(\epsilon) = \tilde{\lambda}_r + O(\epsilon^2)$ as $\epsilon \to 0$, uniformly for $r = 1, 2, \ldots$, with corresponding normalized eigenfunction $\phi_r^-(x,z;\epsilon) = \tilde{A}_r(x) + O(\epsilon^2)$ as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{M}'$.

We can now use the above theory to obtain the following expression for $\tilde{p} : \bar{M}' \times [0,\infty) \mapsto \mathbb{R}$, via (2.46) and (2.47):

$$\tilde{p}(x,z,t) = \sum_{r=1}^{\infty} c_r \mathrm{e}^{-\tilde{\lambda}_r t}\tilde{A}_r(x) + O(\epsilon^2 \mathrm{e}^{-\tilde{\lambda}_1 t}, \mathrm{e}^{-t/\epsilon^2}) \quad \text{as } \epsilon \to 0, \tag{4.20}$$

uniformly for $(x,z,t) \in \bar{M}' \times [\delta,\infty)$, for any $\delta > 0$. Here c_r, $r = 1, 2, \ldots$, are given by

$$c_r = \iint_{\bar{M}'} \tilde{p}_0(u,v)\tilde{A}_r(u)\,\mathrm{d}u\,\mathrm{d}v, \quad r = 1,2,\ldots. \tag{4.21}$$

We observe from (4.20) that

$$\tilde{p}(x,z,t) \sim (c_1\tilde{A}_1(x) + O(\epsilon^2))\mathrm{e}^{-\tilde{\lambda}_1 t} \tag{4.22}$$

as $t \to \infty$, uniformly for $(x,z) \in \bar{M}'$. Thus, the solution to [IBVP] approaches the solution to [SSP] as $t \to \infty$ through terms exponentially small in t as $t \to \infty$. The timescale for relaxation to the steady state is then $t_s \sim (\tilde{\lambda}_1)^{-1}$ in dimensionless variables, giving the dimensional relaxation timescale as $t_s^d \sim c_t l^2/(D_0\tilde{\lambda}_1)$, via (2.12).

5. The case of disparate permeabilities. In the previous sections, the theory has been developed for the situation when the permeability scale is comparable in both the x-direction and the z-direction. In some applications, this is not always the case, when the permeability in the z-direction is much weaker than that in the x-direction. In this case (2.1) should be replaced by

$$D_0^x D_x\left(\frac{x}{l}, \frac{z}{h}\right) \geq D_m > 0, \quad D_0^z D_z\left(\frac{x}{l}, \frac{z}{h}\right) \geq D_m > 0,$$

for $(x,z) \in \bar{M}$, with $D_0^x > 0$ being the permeability scale in the x-direction and $D_0^z > 0$ being the permeability scale in the z-direction, and now

$$0 < \delta = \frac{D_0^z}{D_0^x} \ll 1. \tag{5.1}$$

For such reservoirs δ is typically of $O(10^{-1})$. We now follow the same nondimensionalization as before, via (2.12), with D_0^x replacing D_0. The resulting full dimensionless problem is identical to [IBVP], except with ϵ replaced by $\tilde{\epsilon}$, where $\tilde{\epsilon} = \epsilon\delta^{-1/2}$. Thus all of the previous theory carries over to this situation, on simply replacing ϵ by $\tilde{\epsilon}$. In particular, the asymptotic theory as $\epsilon \to 0$ is now replaced by $\tilde{\epsilon} \to 0$, and so requires

$$0 < \tilde{\epsilon} \ll 1, \tag{5.2}$$

which, with (5.1), is equivalent to $0 < \epsilon \ll \delta^{1/2} \ll 1$, and with typically $\epsilon^2 \sim O(10^{-4})$ and $\delta \sim O(10^{-1})$, then this ordering is satisfied. It is worth noting here that for porous layers in which $\epsilon = O(1)$ and $\delta \gg 1$, then $\tilde{\epsilon} = \epsilon\delta^{-1/2} \ll 1$, so (5.2) is again satisfied, and the asymptotic theory developed before is again applicable.

6. The pseudosteady state. In this section we consider the situation when the specified flux constants α_i, $i = 1, \ldots, N$, do not satisfy the condition (2.31); that is, when $\sum_{i=1}^{N} \alpha_i = \alpha_T \neq 0$. In this case we introduce an associated pseudosteady state problem to [IBVP]. This corresponds to the steady state problem [SSP] ((2.23)–(2.30)), except that now (2.23) is modified to

$$\hat{u}_x + \hat{w}_z = \sum_{i=1}^{N} s_i(z)\delta(x - x_i) - \tilde{\alpha}_T, \quad (x, z) \in M',$$

with the constant $\tilde{\alpha}_T$ given by $\tilde{\alpha}_T = \alpha_T/\text{meas}(\bar{M}')$. The result corresponding to Theorem 2.2 is now, with the pseudosteady state problem referred to as [PSSP], the following.

THEOREM 6.1. *For each $\epsilon > 0$,* [PSSP] *has a unique (up to addition of a constant in $\hat{p}$) solution $\hat{u}, \hat{w}, \hat{p} : \bar{M}' \mapsto \mathbb{R}$.*

Again, we fix the indeterminate constant in [PSSP] to be that pseudosteady state which satisfies the condition (2.40). The result corresponding to Theorem 2.3 is now as follows.

THEOREM 6.2. *For each $\epsilon > 0$,* [IBVP] *has a unique solution $u, w, \bar{p} : \bar{M}' \times [0, \infty) \mapsto \mathbb{R}$ given by*

$$\begin{aligned}\bar{p}(x, z, t) &= \tilde{\alpha}_T t + \hat{p}(x, z) + \tilde{p}(x, z, t),\\ u(x, z, t) &= \hat{u}(x, z) - D_x(x, z)\tilde{p}_x(x, z, t),\\ w(x, z, t) &= \hat{w}(x, z) - \epsilon^{-2}D_z(x, z)\tilde{p}_z(x, z, t),\end{aligned}$$

for all $(x, z) \in \bar{M}'$ and $t \in [0, \infty)$. Here $\tilde{p} : \bar{M}' \times [0, \infty) \mapsto \mathbb{R}$ is given by (2.46), (2.47), *and $\hat{u}, \hat{w}, \hat{p} : \bar{M}' \mapsto \mathbb{R}$ is that solution to* [PSSP] *which satisfies* (2.40). *Moreover,*

$$\begin{aligned}\bar{p}(x, z, t) &= \tilde{\alpha}_T t + \hat{p}(x, z) + O(\mathrm{e}^{-\lambda_1(\epsilon)t}),\\ u(x, z, t) &= \hat{u}(x, z) + O(\mathrm{e}^{-\lambda_1(\epsilon)t}),\\ w(x, z, t) &= \hat{w}(x, z) + O(\mathrm{e}^{-\lambda_1(\epsilon)t}),\end{aligned}$$

as $t \to \infty$, uniformly for $(x, z) \in \bar{M}'$.

We remark that since $\hat{p} \in C(\bar{M}') \cap PC^1(\bar{M}')$ and the initial data for [IBVP] $\bar{p}_0 \in C(\bar{M}') \cap PC^1(\bar{M}')$, then Theorem 6.2 implies global asymptotic stability (up to the addition of a constant to $\hat{p}$) for the pseudosteady state to [IBVP] with respect to perturbations in $C(\bar{M}') \cap PC^1(\bar{M}')$.

We now explore the structure of the solution to [PSSP] as $\epsilon \to 0$. This involves only minor adjustments to the structure developed in section 3 for the solution to

[SSP]. Again the outer region asymptotic expansions are given by

$$\begin{aligned}\hat{u}(x,z;\epsilon) &= \frac{-D_x(x,z)}{\bar{D}_x(x)}\left\{S(x)+\tilde{\alpha}_T H(x)\right\}+O(\epsilon^2),\\ \hat{w}(x,z;\epsilon) &= S(x)\int_{z_-(x)}^{z}\left\{\frac{D_x(x,\lambda)}{\bar{D}_x(x)}\right\}_x \mathrm{d}\lambda\\ &\quad -\frac{z_-'(x)D_x(x,z_-(x))S(x)}{\bar{D}_x(x)}-\tilde{\alpha}_T(z-z_-(x))+O(\epsilon^2),\\ \hat{p}(x,z;\epsilon) &= A(x)+O(\epsilon^2),\end{aligned} \tag{6.1}$$

as $\epsilon \to 0$, uniformly for $(x,z) \in \bar{N}_\epsilon'$. Here, as before, $S : [-1,1] \mapsto \mathbb{R}$ is as given by (3.20), but now $A : [-1,1] \mapsto \mathbb{R}$ is the unique solution to the linear inhomogeneous boundary value problem,

$$[\bar{D}_x(x)A'(x)]' = -\sum_{i=1}^{N}\alpha_i\delta(x-x_i)+\tilde{\alpha}_T(z_+(x)-z_-(x)), \quad x\in(-1,1), \tag{6.2}$$

$$A'(-1)=A'(1)=0, \tag{6.3}$$

$$\int_{-1}^{1}(z_+(x)-z_-(x))A(x)\,\mathrm{d}x = I_0, \tag{6.4}$$

with $H : [-1,1] \mapsto \mathbb{R}$ defined by

$$H(x)=\int_{-1}^{x}(z_+(\lambda)-z_-(\lambda))\,\mathrm{d}\lambda, \quad x\in[-1,1]. \tag{6.5}$$

The solution to (6.2)–(6.4) is readily obtained as

$$A(x)=\int_{-1}^{x}\frac{[S(\lambda)+\tilde{\alpha}_T H(\lambda)]}{\bar{D}_x(\lambda)}\,\mathrm{d}\lambda + A_0, \quad x\in[-1,1], \tag{6.6}$$

with the constant A_0 given by

$$A_0=\frac{I_0}{\mathrm{meas}(\bar{M}')}-\frac{1}{\mathrm{meas}(\bar{M}')}\int_{-1}^{1}\frac{[S(\lambda)+\tilde{\alpha}_T H(\lambda)]}{\bar{D}_x(\lambda)}\,\mathrm{meas}(\bar{M}'(\lambda))\,\mathrm{d}\lambda. \tag{6.7}$$

Thus we now have

$$A_i = A(x_i)=\int_{-1}^{x_i}\frac{[S(\lambda)+\tilde{\alpha}_T H(\lambda)]}{\bar{D}_x(\lambda)}\,\mathrm{d}\lambda + A_0, \tag{6.8}$$

$$A_i^{+'} = A'(x_i^+) = -\frac{\sum_{j=0}^{i}\alpha_j}{\bar{D}_x(x_i)}+\frac{\tilde{\alpha}_T H(x_i)}{\bar{D}_x(x_i)}, \tag{6.9}$$

$$A_i^{-'} = A'(x_i^-) = -\frac{\sum_{j=0}^{i-1}\alpha_j}{\bar{D}_x(x_i)}+\frac{\tilde{\alpha}_T H(x_i)}{\bar{D}_x(x_i)}, \tag{6.10}$$

each for $i = 1,\ldots,N$. The details of the inner regions are precisely as before in section 3, but now A_i and $A_i^{\pm'}$ are given by (6.8)–(6.10). The modifications are now complete.

7. An example. We finally apply the theory developed in the previous sections to a simple situation. We consider a rectangular porous layer, with $z_{\pm}(x) = \pm 1/2$, $x \in [-1,1]$, with the permeability of the layer in the x-direction independent of x, so that $D_x(x,z) = D_x(z)$, $(x,z) \in \bar{M}'$. Thus,

$$\bar{D}_x(x) = \int_{-1/2}^{1/2} D_x(z)\,\mathrm{d}z = \bar{D}_x\ (>0), \quad x \in [-1,1],$$

with $\bar{D}_x$ a constant. We include a single source/sink at $x = x_1 = x_s \in (-1,1)$, with normalized flux constant $\alpha = \alpha_1 = \alpha_s = \pm 1$, with -1 for extraction and $+1$ for injection. We take the initial pressure field to be uniform, so that

$$\bar{p}_0(x,z) = \bar{p}_0, \quad (x,z) \in \bar{M}', \tag{7.1}$$

with $\bar{p}_0$ a constant. The pressure field in the porous layer is then given by

$$\bar{p}(x,z,t;\epsilon) = \frac{1}{2}\alpha_s t + A(x) + c_1 \tilde{A}_1(x)\mathrm{e}^{-\tilde{\lambda}_1 t} + O(\epsilon, \mathrm{e}^{-\tilde{\lambda}_2 t}, \epsilon^2 \mathrm{e}^{-\tilde{\lambda}_1 t}, \mathrm{e}^{-t/\epsilon^2}), \tag{7.2}$$

as $\epsilon \to 0$, uniformly for $(x,z,t) \in \bar{M}' \times [\delta,\infty)$ (for any $\delta > 0$), via Theorem 6.2, together with (6.1), (3.34), (3.41), and (4.20)–(4.22). From (4.17)–(4.19),

$$\tilde{\lambda}_1 = \frac{1}{4}\bar{D}_x \pi^2, \qquad \tilde{\lambda}_2 = \bar{D}_x \pi^2, \qquad \tilde{A}_1(x) = \cos\frac{1}{2}\pi(x+1), \quad x \in [-1,1], \tag{7.3}$$

with, via (4.21), (7.1), and (6.1),

$$c_1 = -\int_{-1}^{1} A(\lambda)\cos\frac{1}{2}\pi(\lambda+1)\,\mathrm{d}\lambda. \tag{7.4}$$

It also follows from (6.2)–(6.7) that

$$A(x) = \begin{cases} \frac{\alpha_s}{4\bar{D}_x}(x+1)^2 + A_0, & x \in [-1, x_s), \\ \frac{\alpha_s}{4\bar{D}_x}(x+1)^2 - \frac{\alpha_s}{\bar{D}_x}(x - x_s) + A_0, & x \in [x_s, 1], \end{cases} \tag{7.5}$$

with $A_0 = \bar{p}_0 - \frac{\alpha_s}{12\bar{D}_x}(1 + 6x_s - 3x_s^2)$. The pressure at the line source/sink is thus

$$\bar{p}(x_s,z,t;\epsilon) = \frac{1}{2}\alpha_s t + A(x_s) + c_1 \tilde{A}_1(x_s)\mathrm{e}^{-\tilde{\lambda}_1 t} + O(\epsilon, \mathrm{e}^{-\tilde{\lambda}_2 t}, \epsilon^2 \mathrm{e}^{-\tilde{\lambda}_1 t}, \mathrm{e}^{-t/\epsilon^2}),$$

from which it follows, via (7.3)–(7.5), that

$$\begin{aligned} \bar{p}(x_s,z,t;\epsilon) = \frac{1}{2}\alpha_s t + \bar{p}_0 + \frac{\alpha_s}{6\bar{D}_x}(1+3x_s^2) + c_1 \mathrm{e}^{-\frac{\bar{D}_x\pi^2}{4}t}\cos\frac{1}{2}\pi(x_s+1) \\ + O\left(\epsilon, \mathrm{e}^{-\bar{D}_x\pi^2 t}, \epsilon^2 \mathrm{e}^{-\frac{\bar{D}_x\pi^2}{4}t}, \mathrm{e}^{-t/\epsilon^2}\right) \end{aligned} \tag{7.6}$$

as $\epsilon \to 0$, uniformly for $(z,t) \in [-1/2,1/2] \times [\delta,\infty)$. To obtain the correction at $O(\epsilon)$ to (7.2), we note that a composite expansion must first be obtained using the inner and outer expansions for $\bar{p}$, but this is not necessary at leading order. As $t \to \infty$, (7.6) gives

$$\bar{p}(x_s,z,t;\epsilon) \sim \bar{p}_0 + \frac{1}{2}\alpha_s\left[t + \frac{1}{3\bar{D}_x}(1+3x_s^2)\right]. \tag{7.7}$$

Finally, denoting the dimensionless atmospheric pressure by $\bar{p}_a$, then with the initial layer pressure $\bar{p}_0 \gg \bar{p}_a$, we may use (7.7) to obtain the time span during which the single well is self-producing at the specified extraction rate. With an extraction well, $\alpha_s = -1$ in (7.7), and the time limit of self-production is $t = t_c$, where

$$\bar{p}(x_s, z, t_c; \epsilon) = \bar{p}_a. \tag{7.8}$$

On using (7.7) in (7.8), we obtain

$$t_c = 2(\bar{p}_0 - \bar{p}_a) - \frac{1}{3\bar{D}_x}(1 + 3x_s^2). \tag{7.9}$$

It follows from (7.9) that t_c is optimized by locating the extraction well at $x = x_s = 0$, that is, at the center of the porous layer as should be expected due to the symmetry in this simple example. The point is that, in less symmetrical examples, optimization can be achieved with little more effort through the corresponding version of (7.9), which is still readily available. In dimensional terms, (7.9) becomes, via (2.12),

$$t_c^d = \frac{ac_t}{Q}(p_0^d - p_a) - \frac{hc_t}{3\bar{D}_x^d}(l^2 + 3x_s^{d^2}),$$

with t_c^d the dimensional self-extraction time, $a = 2hl$ the cross-sectional area of the porous layer, Q the volumetric extraction rate per unit width, p_0^d the dimensional initial layer pressure, p_a the dimensional atmospheric pressure, $\bar{D}_x^d$ the dimensional depth integrated permeability in the x-direction, and x_s^d the dimensional location of the extraction well.

8. Conclusions.

In this paper we have considered the unsteady flow of a weakly compressible fluid in a horizontal layer of an inhomogeneous and anisotropic porous medium with variable upper and lower boundaries, in the presence of line sources and sinks. We have derived a strongly parabolic linear initial boundary value problem for the dynamic fluid pressure and shown that this problem has a unique solution. We have then constructed the solution to this problem when the layer aspect ratio $0 < \epsilon \ll 1$, via the method of matched asymptotic expansions. First, we have derived a matched asymptotic solution to the steady state problem, under the constraint that the sum of the total volume fluxes at the wells is zero. (This constraint is removed in section 6, leading to a pseudosteady state problem whose solution is almost identical in structure.) In the outer region this has been constructed directly, with the solution given by (3.30)–(3.32). In the inner region the solution is given by (3.63)–(3.66), together with (3.34) and (3.41). This solution is written in terms of the eigenvalues and eigenvectors of a regular Sturm–Liouville eigenvalue problem [SL], which can be solved analytically in the case that the permeability at each line source/sink is constant in the vertical direction, but whose numerical solution is straightforward in the more general case. The pressure at any line source or sink is then given by (3.67).

By subtracting the solution of the steady state problem from the solution of the initial value problem, we have then constructed a strongly parabolic homogeneous problem with no discontinuities across the line sources and sinks, whose solution can be written in terms of the eigenvalues and eigenfunctions of a regular self-adjoint eigenvalue problem. Asymptotic solution of this reduces to solution of a regular Sturm–Liouville eigenvalue problem identical in structure to [SL]. It has further been shown, via (4.20)–(4.22), that the solution of the initial value problem approaches the solution of the steady state problem through terms exponentially small with respect to time t as $t \to \infty$. Generalizations to cases where a line source or sink is near a boundary wall, where line sources and sinks are not well spaced, and to the case of disparate permeabilities have also been considered, in sections 3.1, 3.2, and 5, respectively. An example demonstrating an application of the theory to a simple situation is provided in section 7.

We finally remark that since the initial boundary value problem is solved for a general C^1 initial condition, the effect of time dependent transient effects due to temporal changes in the well discharge rates can easily be accounted for.

REFERENCES

[1] A.-J. A. Al-Khalifah, K. Aziz, and R. N. Horne, *A new approach to multiphase well test analysis*, in Society of Petroleum Engineers 16743, SPE, Houston, TX, 1987.

[2] K. Aziz and A. Settari, *Petroleum Reservoir Simulation*, Applied Science Publishers, London, 1979.

[3] D. K. Babu and A. S. Odeh, *Productivity of a horizontal well*, in Society of Petroleum Engineers 18334, SPE, Houston, TX, 1988.

[4] G. S. Busswell, R. Banerjee, R. K. M. Thambynayagam, and J. B. Spath, *Generalized analytical solution for reservoir problems with multiple wells and boundary conditions*, in Society of Petroleum Engineers 99288, SPE, Houston, TX, 2006.

[5] H. Cinco-Ley, F. Samaniego-V, and A. N. Dominguez, *Transient pressure behavior for a well with a finite-conductivity vertical fracture*, Soc. Petroleum Engineers J., 18 (1978), pp. 253–264.

[6] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, Toronto, London, 1955.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[8] J. P. Gilchrist, G. S. Busswell, R. Banerjee, J. B. Spath, and R. K. M. Thambynayagam, *Semi-analytical solution for multiple layer problems with multiple vertical, horizontal, deviated and fractured wells*, in Proceedings of the International Petroleum Technology Conference, Dubai, U.A.E., 2007, SPE, Houston, TX, paper 11718.

[9] E. Gomes and Z. A. Reza, *A new semi-analytic pressure transient model for layered gas reservoir under various reservoir and well conditions*, in Society of Petroleum Engineers 39526, SPE, Houston, TX, 1998.

[10] A. Gringarten and H. J. Ramey, Jr., *The use of source and Green's functions in solving unsteady-flow problems in reservoirs*, Soc. Petroleum Engineers J., 13 (1973), pp. 285–296.

[11] D. J. Needham and S. Langdon, *The Unsteady Flow of a Weakly Compressible Fluid in a Thin Porous Layer.* II: *Three-dimensional Theory*, Mathematics Department Preprint Series MPS_2009_01, University of Reading, UK.

[12] M. Oguztoreli and D. W. Wong, *Vertex: A new modeling method to direct field development*, in Society of Petroleum Engineers 39806, SPE, Houston, TX, 1998.

[13] E. Ozkhan and R. Raghavan, *Well test analysis problems: Part 1—Analytical considerations*, in Society of Petroleum Engineers 18615, SPE, Houston, TX, 1991.

[14] R. Pecher, S. D. Harris, R. J. Knipe, L. Elliot, and D. B. Ingham, *New formulation of the Green element method to maintain its second order accuracy in 2D/3D*, Engineering Analysis with Boundary Elements, 25 (2001), pp. 211–219.

[15] R. Raghavan, *Well test analysis: Wells producing by solution gas drive*, in Society of Petroleum Engineers 5588, SPE, Houston, TX, 1975.

[16] A. G. Ramm, *Limit spectra of the interior Neumann problems when a solid domain shrinks to a plane one*, J. Math. Anal. Appl., 108 (1985), pp. 107–112.

[17] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

[18] R. K. M. Thambynayagam, *Diffusion: A Compendium of Analytical Solutions*, in preparation.

[19] L. G. Thompson, J. L. Manrique, and T. A. Jelmeri, *Efficient algorithms for computing the bounded reservoir horizontal well pressure response*, in Society of Petroleum Engineers 21827, SPE, Houston, TX, 1991.

[20] M. Van Dyke, *Perturbation Methods in Fluid Mechanics*, The Parabolic Press, Stanford, CA, 1975.

SIAM J. APPL. MATH.
Vol. 69, No. 4, pp. 1110–1161

GROWTH DYNAMICS OF CELL ASSEMBLIES*

MAU-HSIANG SHIH† AND FENG-SHENG TSAI†

Abstract. We explore an evolutionary network model of pulse-coupled neurons in which the changes of evolutionary coupling strengths are based on Hebbian synaptic plasticity. We show that the ongoing changes of the evolutionary network's nodal-and-coupling dynamics will eventually result in group synchrony and sync-dependent circuits. We also tackle the problem of the stability of neural synchrony and the problem of determining the size of synchronously firing neural groups. This leads to describing a phenomenon underlying synchrony and stability of synchrony that neural synchrony allows positive feedback from which a monotonically increasing sequence of coupling strengths and a monotonically increasing region of states for initializing the stability process arise.

Key words. synaptic plasticity, cell assemblies, synchronization, stability, self-organization, spontaneous order, complex networks, nonlinear dynamics

AMS subject classifications. 37F20, 92B20, 00A71, 68T05, 91E40

DOI. 10.1137/070697471

1. Introduction.

The brain is considered to be a complex, self-organizing system. It consists of enormous numbers of interacting neurons and perpetually weaves its intricate web [7, 21, 42, 50, 52]. What underlies such activation is plasticity; it emerges as a source of change altering the structure of the brain. In 1949, Donald O. Hebb proposed an activity-dependent mechanism for synaptic modification, which was the first neurophysiological description in what is now called "Hebbian synaptic plasticity." It rested on his most famous statement: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." Synapses that exhibit this coincidence-detection rule are now called "Hebb synapses" [21, 41]. Hebb went further and suggested that cortical circuits might admit self-sustaining reverberatory activity to bind association-area neurons into structural interconnectivity if the changes of synaptic strengths could be based on coincidence detection [21, 41]. The hypothesis is now called the "Hebbian cell-assembly postulate," which is one of the most influential postulates in relation to the behavior and neural interactions in the brain [1, 21]. According to the Hebbian cell-assembly postulate, the circulating neural impulses between populations of association-area neurons would continue to circulate, forming a diffuse self-assembling structure called "cell assemblies" [21, 36, 41].

Being on the trail of the growth of cell assemblies always accompanies complicated integration of neuronal and synaptic activity [20, 34, 38]. It stimulates an intensive effort to promote the building of computer or network models of the brain [2, 3, 17, 23, 57] and leads to a shift towards explaining cognitive function in terms of large-scale interactions of neuronal populations [11, 13, 15, 16, 40, 47]. This development may lead to the current research on self-organization, which probes the interplay between neural activity and synaptic plasticity. It lends substance to the formation

*Received by the editors July 17, 2007; accepted for publication (in revised form) July 9, 2008; published electronically January 30, 2009. This research was supported in part by the National Science Council of the Republic of China.

http://www.siam.org/journals/siap/69-4/69747.html

†Department of Mathematics, National Taiwan Normal University, 88 Sec. 4, Ting Chou Road, Taipei 116, Taiwan (mhshih@math.ntnu.edu.tw, fstsai@abel.math.ntnu.edu.tw).

of connectivity structures in cortical development and gives birth to abstract neural network models incorporated with dynamical and structural complexity [18, 25, 29, 30, 31, 54, 56].

Self-organization may proceed on a layer-by-layer basis, in which the synaptic strengths are repeatedly modified in response to input patterns and in accordance with the rules underlying plasticity. Willshaw and Malsburg in 1976 proposed a mathematical topography formation model which was capable of forming topographic connections between two layers [56]. Also, Kohonen in 1982 proposed the self-organizing map (SOM) algorithm, which describes a competitive learning rule to form a topology preserving mapping [27, 28, 29]. An almost-sure convergence of the SOM algorithm was given by Forte and Pagés [14]. Additionally, within the next few years, various extensions to the SOM algorithm were proposed. Examples include the temporal Kohonen map [8], which extends the SOM by adding the activity of leaky integrators to the SOM, and the recursive SOM [19, 53], which extends the SOM by establishing feedback to represent time. More recently, Lücke and Malsburg have discussed a self-organization process which reflects the hierarchical nature of receptive field formation [30, 31]. They take a minicolumn to consist of excitatory McCulloch–Pitts neurons and show that if the dynamics is weakly coupled to input by afferent fibers and subjects to Hebbian synaptic plasticity, then a self-organization of minicolumnar receptive fields is induced.

Alternatively, self-organization may proceed on a recurrent network. Hopfield in 1982 initiated a recurrent network of nerve cells, whose couplings are established in response to input patterns and in accordance with Hebbian synaptic plasticity [23]. He used the approach of energy minimization to show that the network's dynamics will tend toward a stable equilibrium state when the retrieval operation is performed asynchronously. The Hopfield network was designed to work as a content-addressable memory (CAM) on the basis of collective dynamics and computing with attractors. Meanwhile, Cohen and Grossberg in 1983 initiated a general model of a nonlinear cooperative-competitive neural network [9, 17]. Cohen and Grossberg constructed a global Lyapunov function for assessing the stability of network dynamics and described a general principle for designing CAM networks.

However, the previous models for self-organization are inherently static, with time taking a secondary role in network architecture [37]. These models require preprocessors to encode input patterns in synaptic weight matrices, thereby converting temporal dynamic information into static spatial information outside the network. Hebbian synaptic plasticity is used in the construction of a synaptic weight matrix only for the initial input patterns but not for network evolution.

This motivates us to study self-organization by proceeding on a neural network entwined with nonlinearity and dynamism. What we need is the quest for a deep theory of complex networks, which allows for describing structural complexity, network evolution, dynamical complexity, and meta-complication [49]. We build a model of an evolutionary network consisting of enormous numbers of McCulloch–Pitts neurons, each simple, but myriad interactions between them could be extremely complicated. The influence of McCulloch–Pitts neurons was very much in the thoughts of von Neumann as he developed his ideas for the modern digital computer [5, 17], and was very much in the work of Minsky in automata theory and theory of computation [35].

In our model of the evolutionary network, the time- and activity-dependent nodal-and-coupling changes are based on an algorithmic aspect of Hebbian synaptic plasticity. Each change reflects the interplay and large-scale integration between neuronal and synaptic activity. We take the view that Hebb synapses are crucial to the generation of neural synchrony [46, 48, 52, 58] and to the development of diffuse structure of cell assemblies [20, 36, 41]. This leads to addressing the mathematical problems of the tendency of neurons to synchronize underlying Hebbian synaptic plasticity, the stability of neural synchrony, the effect of neural synchrony, and the determination of the size of synchronously firing neural groups.

2. Evolutionary network model. According to the all-or-none character of neural activity, McCulloch and Pitts in 1943 introduced a binary processing unit which performed simple threshold logic. They pointed out that the brain was potentially a powerful logic and computational device [3, 33]. Here, through the use of McCulloch–Pitts neurons, we will construct an evolutionary network which straightens out an alternating change in its nodal and coupling dynamics. The model of the evolutionary network we are concerned with consists of a population of n distinct integrate-and-fire processing units (McCulloch–Pitts neurons or neurons) [33, 35]; each constantly integrates all incoming signals transferred from synapses on its cell body and dendrites, and fires action potentials to send signals to other neurons when the combined effect reaches its threshold. Name those *neurons* $1, \dots, n$ and denote by the ordered pair (i, j) the *evolutionary coupling* linking neuron j to neuron i. All the evolutionary couplings are fundamentally dynamic, be they symmetrical [23, 57] or unsymmetrical, hierarchical, small-world [54], or reverberating-circuit [44], to reflect the interconnected neurons in the brain. To each neuron i there is associated the *threshold* b_i and the *active state variable* $x_i = 0$ or 1, and to each evolutionary coupling (i, j) there is associated the *coupling strength variable* a_{ij}. The phase space of the evolutionary network of n coupled neurons is denoted by $\{0, 1\}^n$, the binary code consisting of all 01-strings $x_1x_2 \cdots x_n$ of fixed-length n.

Fix $t = 0, 1, \dots$ for the moment. The corresponding *neuronal active state* and *evolutionary coupling state* are denoted by $x(t) = (x_1(t), x_2(t), \dots, x_n(t))$ and $A(t) = [a_{ij}(t)]_{n\times n}$, respectively. The function *hea* is the Heaviside function: $hea(u) = 1$ for $u \geq 0$, otherwise 0, which describes an instantaneous unit pulse. To generate the neuronal active state $x(t + 1)$ and the evolutionary coupling state $A(t + 1)$, we have to introduce the function $H_{A(t),s(t)}$ and the plasticity parameter $\mathcal{D}_{x(t)\to x(t+1)} a_{ij}$. We associate to t a nonempty subset $s(t)$ of $\{1, 2, \dots, n\}$ (denoting the neurons that adjust their activity at time t) and a function $H_{A(t),s(t)} : \{0, 1\}^n \longrightarrow \{0, 1\}^n$ whose ith component is defined by

$$[H_{A(t),s(t)}(x)]_i = x_i \quad \text{if } i \notin s(t),$$

otherwise

$$[H_{A(t),s(t)}(x)]_i = hea\left(\sum_{j=1}^{n} a_{ij}(t)x_j - b_i\right),$$

such that

$$x(t + 1) = H_{A(t),s(t)}(x(t)). \tag{1}$$

For every $i, j = 1, 2 \dots, n$, denote by $\mathcal{D}_{x(t)\to x(t+1)} a_{ij}$ the parameter which is a representative for a choice of real numbers, so that the evolutionary coupling state $a_{ij}(t + 1)$ at (i, j) is given by the parametric equation

$$a_{ij}(t + 1) = a_{ij}(t) + \mathcal{D}_{x(t)\to x(t+1)} a_{ij}. \tag{2}$$

When $\mathcal{D}_{x(t)\to x(t+1)}a_{ij}$ varies, $a_{ij}(t+1)$ changes. And this is how plasticity is created in the dynamics of the evolutionary network. The parameter $\mathcal{D}_{x(t)\to x(t+1)}a_{ij}$ in (2) is called the *plasticity parameter* of the evolutionary coupling (i,j), which varies with respect to the neuronal active state changing from $x(t)$ to $x(t+1)$. Now let t vary. The alternating nature of (1) and (2) reveals a dynamical-combinatorial process in which the neuronal active state and the evolutionary coupling state keep changing, looping back on one another with extremely fast switches (at the millisecond level) and giving rise to patterns of indescribable complexity. The neuronal active state changing from $x(t)$ to $x(t+1)$ by (1) leads to the choices of plasticity parameters and results in the changes of the evolutionary coupling state from $a_{ij}(t)$ to $a_{ij}(t+1)$ by (2). The changes of the evolutionary coupling state loop back on the changes of the neuronal active state from $x(t+1)$ to $x(t+2)$ by (1), and then on the changes of the evolutionary coupling state from $a_{ij}(t+1)$ to $a_{ij}(t+2)$ by (2), and continue recursively. So we have a nonlinear dynamical system of the n coupled neurons modeled by the following nonlinear parametric equations:

$$x(t+1) = H_{A(t),s(t)}(x(t)), \quad t = 0,1,\ldots, \tag{3}$$

$$A(t+1) = A(t) + D_{x(t)\to x(t+1)}A, \quad t = 0,1,\ldots, \tag{4}$$

where $H_{A(t),s(t)}(x)$ are the time-and-state varying functions encoding the dynamics, and each $\mathcal{D}_{x(t)\to x(t+1)}A$ is an n-by-n real matrix whose (i,j)-entry is $\mathcal{D}_{x(t)\to x(t+1)}a_{ij}$. The plasticity parameters quantify plasticity of evolutionary couplings that allows the system as a whole to undergo spontaneous organization.

There are many different ways in which we carry out the updating specified by the choice of $s(t)$ for $t = 0,1,\ldots$. Let us call that the discrete flow $x(t)$ generated by (3) and (4) *iterates asynchronously* if $s(t)$ is a singleton for all $t = 0,1,\ldots$ and the union $\cup_{t\geq\tau}s(t)$ equals $\{1,2,\ldots,n\}$ for any $\tau \geq 0$. We can begin asynchronous updating in the way that we select at random a neuron to adjust its activity at each time step $t = 0,1,\ldots,$ or equivalently, from an autonomous point of view, we can begin asynchronous updating in the way that each neuron independently chooses to adjust its activity with some constant probability per unit time [22]. The latter can generate a random sequence of updating neurons in time because there is vanishingly small probability of two neurons choosing to adjust themselves at exactly the same moment. Here we adopt asynchronous updating which allows us to concentrate on the alternating changes between the nodal and coupling dynamics.

Let us note that based on asynchronous updating, we can generate a specific evolutionary network whose dynamics obey the Gauss–Seidel iteration. To see this, consider the case $s(t) = (t/n - [t/n])n + 1$, where $[t/n]$ denotes the greatest integer less than or equal to t/n for $t = 0,1,\ldots,$ and let $x(t)$ and $A(t)$ denote the corresponding neuronal activity state and evolutionary coupling state generated by (3) and (4), respectively. Put $y(t) = x(tn)$ and $w_{ij}(t) = a_{ij}(tn+i-1)$ for every $i,j = 1,2,\ldots,n$ and $t = 0,1,\ldots$. Then we obtain an evolutionary network whose coupling architecture at time t can be defined by the matrix $[w_{ij}(t)]_{n\times n}$, with the sequence of neuronal activity states $\{y(t);\ t = 0,1,\ldots\}$ encoding the network's dynamics. In this evolutionary network, the updating of y_i fulfills the Gauss–Seidel iteration

$$y_i(t+1) = hea\left(\sum_{j=1}^{i-1} w_{ij}(t)y_j(t+1) + \sum_{j=i}^{n} w_{ij}(t)y_j(t) - b_i\right).$$

3. Coincidence-detection evolving algorithm and synchronization problem. In what follows we introduce the coincidence-detection evolving algorithm which provides a way for the choices of plasticity parameters. This algorithm represents a generalized learning rule analogous to the coincidence-detection rule of Hebbian synaptic plasticity. Synaptic modification determined by the coincidence between pre- and postsynaptic activity is used as the bridging mechanism that guides the alternating changes of the evolutionary network's nodal and coupling dynamics.

For every $i, j = 1, 2, \ldots, n$ we first define the *indicator* δ_{ij} on $\{0, 1, \ldots\}$ as follows:

(i) Put $\delta_{ij}(0) = 0$.

(ii) Given $t = 0, 1, \ldots$ and the neuronal active state changing from $x(t)$ to $x(t+1)$ according to the dynamics (3).
If $(x_i(t), x_j(t)) = (1, 1)$ and $(x_i(t+1), x_j(t+1)) = (0, 1)$, put $\delta_{ij}(t+1) = 1$.
If $(x_i(t), x_j(t)) = (0, 1)$ and $(x_i(t+1), x_j(t+1)) = (1, 1)$, put $\delta_{ij}(t+1) = 1$.
If $(x_i(t), x_j(t)) = (1, 1)$ and $(x_i(t+1), x_j(t+1)) = (1, 1)$, put $\delta_{ij}(t+1) = \delta_{ij}(t)$.
If $(x_i(t), x_j(t)) = (0, 1)$ and $(x_i(t+1), x_j(t+1)) = (0, 1)$, put $\delta_{ij}(t+1) = \delta_{ij}(t)$.

(iii) If the pair of $(x_i(t), x_j(t))$ and $(x_i(t+1), x_j(t+1))$ is not in the case of (ii), put $\delta_{ij}(t+1) = 0$.

The binary value $\delta_{ij}(t+1)$ signifies whether the active state of neuron j at time t has a tendency to change the active state of neuron i at time $t+1$. Armed with the indicator δ_{ij}, we define now the *coincidence-detection evolving algorithm*:

For every $t = 0, 1, \ldots$ and $1 \leq i, j \leq n$,

(I) $\mathcal{D}_{x(t)\to x(t+1)} a_{ij} \geq 0$ if $i, j \in \mathbf{1}(x(t+1))$; otherwise $\mathcal{D}_{x(t)\to x(t+1)} a_{ij} \leq 0$;

(II) if $\delta_{ij}(t+1) > \delta_{ji}(t+1)$, then $|\mathcal{D}_{x(t)\to x(t+1)} a_{ij}| \geq |\mathcal{D}_{x(t)\to x(t+1)} a_{ji}|$.

Rule (I) suggests that if two neurons are not active synchronously, the coupling strength between two neurons can be either unchanged or weakened. To be more specific, if the coupling strength is selectively to be unchanged, then rule (I) reduces to an algorithmic aspect of Hebbian synaptic plasticity.

The algorithm basically describes the tendency for changing the evolutionary coupling strengths from $a_{ij}(t)$ to $a_{ij}(t+1)$, which might keep the active (resp., quiescent) neurons active (resp., quiescent) from time $t+1$ to $t+2$. So we obtain, when t varies, myriad groups of neurons (which allow overlapping communities) that are prone to be active or quiescent transiently, dedicating to drive themselves into sustained activity of synchrony. The evolving time the dynamics involves determines the role the coincidence-detection rule plays in accumulating vast numbers of time- and activity-dependent changes in evolutionary couplings, in places forming uncertainty that can grow in the diversity of network evolution (see Figure 1).

Undergoing the dynamical-combinatorial process of network evolution, the neuronal active state $x(t)$ and the evolutionary coupling state $A(t)$ can be generated by (3) and (4), respectively, and we say that neurons in a subset V of $\{1, \ldots, n\}$ are *synchronized* with respect to $x(t)$ if there is a $T \geq 0$ such that the condition $\mathbf{1}(x(t)) = V$ holds true for all $t \geq T$, where $\mathbf{1}(x(t))$ denotes the collection of all active neurons at time t. In the present paper we wish to explore the synchronization problem: *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Do there exist a finite $T \geq 0$ and a subset V of $\{1, \ldots, n\}$ such that $\mathbf{1}(x(t)) = V$ for all $t \geq T$?* To be more specific, the synchronization problem is the mathematical equivalent of the separation problem (see Figure 2): *Do there exist a finite $T \geq 0$ and a subset V of $\{1, \ldots, n\}$*

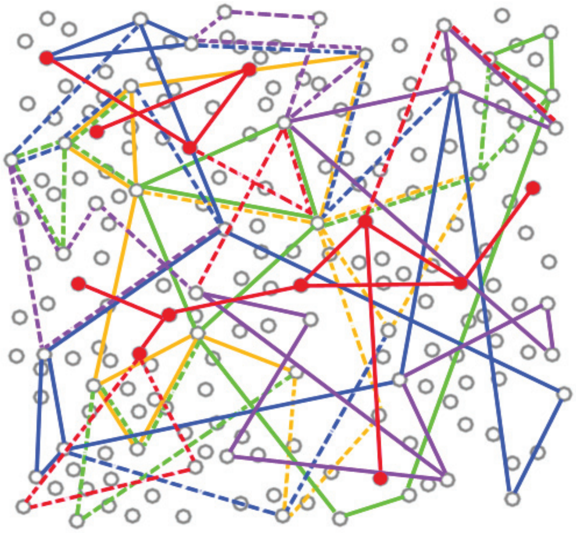


FIG. 1. *The coupling changes based on the coincidence-detection evolving algorithm. Fix t and consider the corresponding active neurons (red nodes). It follows from the coincidence-detection evolving algorithm that the evolutionary couplings will be strengthened (solid red lines) if they are within the group of active neurons (red nodes); otherwise the evolutionary couplings will be weakened (dashed red lines). The nodal dynamics lead to myriad groups of active neurons which can bring about cumulative changes in the shape of interconnectedness. Colors indicate different changes of evolutionary couplings on time steps* $t = 0, 1, \ldots$ *(solid lines: positive plasticity parameters; dashed lines: negative plasticity parameters).*

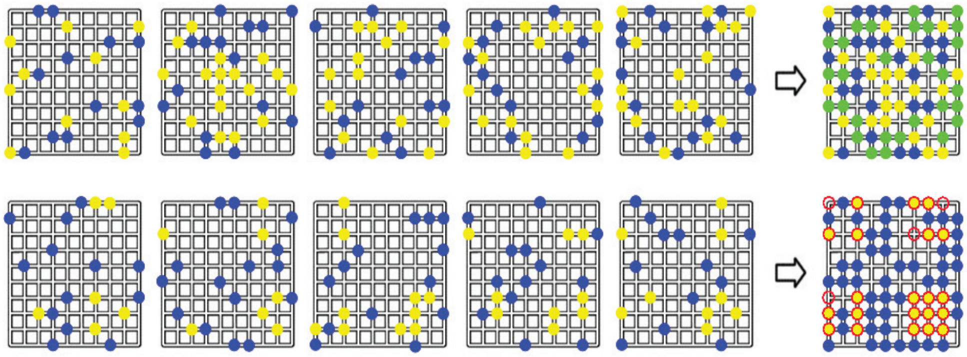


FIG. 2. *The spatial distributions of the positive and negative plasticity parameters. The positive plasticity parameters are represented geometrically as yellow grid points on the* n*-by-*n *grids (for clarity we put* $n = 10$*), and the negative plasticity parameters as blue ones. As the evolutionary network evolves, the spatial distributions of the positive and negative plasticity parameters change from left to right, and the grids at the end of the two rows show the cumulative changes of these distributions (yellow: positive; blue: negative; green: both positive and negative have occurred). The grids in the first row display bewildering patterns, in which the yellow and blue grid points are located in a nearly random way, whereas the grids in the second row exhibit a spontaneously organized process, in which the yellow and blue grid points are located in two fixed clusters separated by red circles (the evolutionary couplings relating to a group of synchronized firing neurons). The separation of the grid points exhibits self-sustaining activity of coupling strengthening when neurons fire in synchrony.*

such that

$$\bigcap_{t\geq T}\{(i,j);\ \mathcal{D}_{x(t)\to x(t+1)}a_{ij}\geq 0\}\supset V\times V\supset\bigcup_{t\geq T}\{(i,j);\ \mathcal{D}_{x(t)\to x(t+1)}a_{ij}>0\}?$$

We notice that previous models of network dynamics have focused on a logical strategy necessary for networked systems to optimize some aspect of their performance, such as using a gradient descent method to minimize the mean square error (the LMS algorithm [55] or the backpropagation algorithm [39]) or using a global Lyapunov method to provide absolute stability of global pattern formation (see [23] for the discrete time Lyapunov function, [9] for an explicit construction of a global Lyapunov function given by Cohen and Grossberg, and [17] for a survey of the Cohen–Grossberg model and its relations to a number of popular models of content-addressable memory). By contrast, the coincidence-detection evolving algorithm we describe here offers predictions in measurable coupling changes merely based on the coincidence-detection rule of Hebbian synaptic plasticity. It reveals that many intricate, outwardly indecipherable patterns of organization can be determined by small changes of coupling strengths depending on the nodal dynamics. Plasticity in evolutionary couplings implies a degree of uncertainty in the dynamics of networked systems, which need not navigate to arrive at a local minimum or maximum of a performance function. Related lines of research in switched linear networked systems have also provided a concise theoretical framework similar to this point [24, 43]. It has shown that in a switched linear networked system all switching sequences of coupled matrices can be asymptotically stable [43, 45], but no common quadratic Lyapunov function exists through the use of a theoretical result of optimal joint spectral radius range for the simultaneous contractibility of coupled matrices [4]. Finding a common quadratic Lyapunov function in such a switched linear networked system becomes increasingly hard as the complexity of coupled matrices and the dimension of the networked system go up [4]. The above discussion suggests that the use of a Lyapunov function might have its own limitations to study the synchronization problem formulated in a high-dimensional nonlinear evolutionary networked system.

4. Driving forces. To solve the synchronization problem, we introduce two evolutionary quantities to measure the driving forces of the evolutionary network's dynamics. We consider the "driving forces" derived from the evolutionary network's nodal and coupling activity, without invoking any Lyapunov function or "physical energy" to control system dynamics.

For any 01-string $x=x_1x_2\cdots x_n$ we define

$$\mathbf{1}(x)=\{i;\ x_i=1,\ 1\leq i\leq n\}$$

and

$$\mathbf{0}(x)=\{i;\ x_i=0,\ 1\leq i\leq n\}.$$

Denote by $\langle\cdot,\cdot\rangle$ the usual scalar product in $\mathbb{R}^n$. Given any two subsets U and V of $\{1,2,\dots,n\}$ and any $t=0,1,\dots$ we define

$$l_t(U,V)=\langle A(t)u,v\rangle,$$

where $u,v\in\{0,1\}^n$ with $\mathbf{1}(u)=U$ and $\mathbf{1}(v)=V$. For every $t=0,1,\dots,$ let

$$[x(t)]^+=\mathbf{0}(x(t))\cap\mathbf{1}(x(t+1))$$

and

$$[x(t)]^- = \mathbf{1}(x(t)) \cap \mathbf{0}(x(t+1)).$$

The *fired-driven strength evaluated at time* t, denoted as $FS(t)$, and the *unfired-driven strength evaluated at time* t, denoted as $US(t)$, are defined by

$$FS(t) = l_t(\mathbf{1}(x(t)), [x(t)]^+)$$

and

$$US(t) = l_t(\mathbf{1}(x(t)), [x(t)]^-).$$

The fired-driven strength $FS(t)$ is not necessarily greater than the unfired-driven strength $US(t)$ over time, but when the discrete flow $x(t)$ behaves in the way that $x(t_*) = x(t^*) \neq x(\hat{t})$ with $t_* < \hat{t} < t^*$ (a feedback loop initiated by active neurons at time t_*), the fired-driven strengths and the unfired-driven strengths in the period of t_* and t^* emerge the orderliness. The order comes from the combined effect of the structural and dynamical complexity of the evolutionary network, in which the updating neuron $s(t)$ and the plasticity parameter $\mathcal{D}_{x(t)\to x(t+1)} a_{ij}$ are arbitrarily chosen for all $t = 0, 1, \dots$ and $i, j = 1, 2, \dots, n$ (see Figure 3).

THEOREM 1. *If $x(t)$ iterates with $x(t_*) = x(t^*) \neq x(\hat{t})$ for some $t_* < \hat{t} < t^*$, then the orderliness*

$$\begin{aligned} FS(t_*) + &FS(t_*+1) + \cdots + FS(t^*-1) \\ &> US(t_*) + US(t_*+1) + \cdots + US(t^*-1) \end{aligned} \tag{5}$$

emerges.

Proof. Let

$$\Lambda^+ = \{t;\ [x(t)]^+ \neq \emptyset,\ t_* \le t < t^*\}$$

and

$$\Lambda^- = \{t;\ [x(t)]^- \neq \emptyset,\ t_* \le t < t^*\}.$$

Then $\Lambda^+ \neq \emptyset$ and $\Lambda^- \neq \emptyset$. Indeed, if $\Lambda^+ = \emptyset$ or $\Lambda^- = \emptyset$, then either

$$\mathbf{1}(x(t_*)) \supset \mathbf{1}(x(t_*+1)) \supset \cdots \supset \mathbf{1}(x(t^*)) \tag{6}$$

or

$$\mathbf{1}(x(t_*)) \subset \mathbf{1}(x(t_*+1)) \subset \cdots \subset \mathbf{1}(x(t^*)). \tag{7}$$

Either (6) or (7) with the condition $x(t_*) = x(t^*)$ gives

$$x(t_*) = x(t_*+1) = \cdots = x(\hat{t}) = \cdots = x(t^*-1) = x(t^*),$$

contradicting the assumption $x(t^*) \neq x(\hat{t})$. The dynamics (3) ensures that

$$[x(t)]^+, [x(t)]^- \subset s(t) \quad \text{for each } t = 0, 1, \dots,$$

$$FS(t) \geq \sum_{j \in [x(t)]^+} b_j \quad \text{for each } t \in \Lambda^+,$$

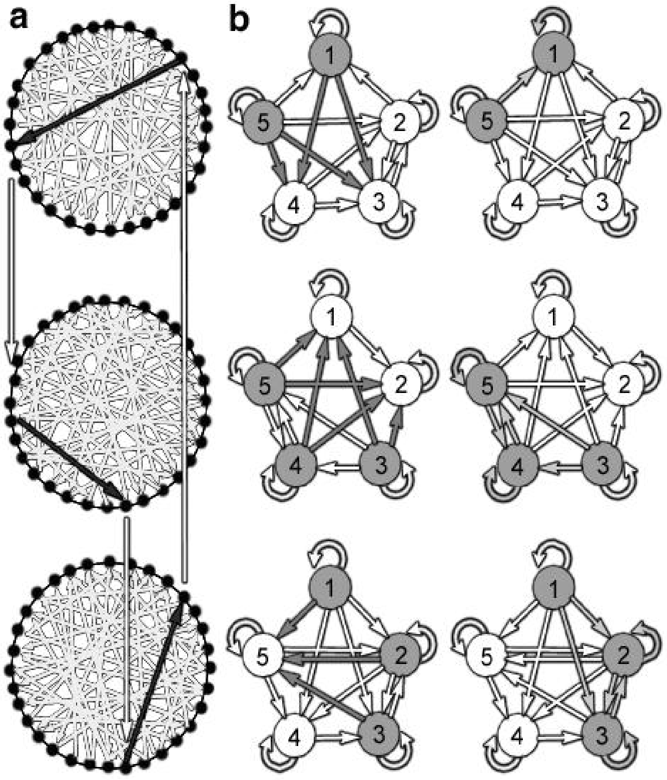


Fig. 3. *The driving forces underlying the evolutionary network's dynamics.* (a) *Each of the three wiring diagrams starts with a ring lattice of* 2^n *nodes (for clarity* $n = 5$ *is shown here); each node denotes one of the neuronal active states in the phase space* $\{0,1\}^n$. *Links (arrows) in each diagram wired from one node* i *to another node* j *indicate the possibility of the changes of neuronal active states from node* i *(at time* t*) to node* j *(at time* $t+1$*). There are many different links starting from each node because of the choices of plasticity parameters. The black arrows show a feedback loop* $x(t) = 10001$, $x(t+1) = 00111$, $x(t+2) = 11100$, *and* $x(t+3) = 10001$ *of one possible discrete flow in the phase space* $\{0,1\}^n$. (b) *The active neurons at time* t, $t+1$, *and* $t+2$, *as seen in* (a), *are indicated by gray nodes in each diagram (from top to bottom). The gray arrows in the left column give a specific way of determining the fired-driven strengths that make quiescent neurons fired at the next time (*t*: neurons* 3 *and* 4*;* $t+1$*: neurons* 1 *and* 2*;* $t+2$*: neuron* 5*), and the gray arrows in the right column illustrate the unfired-driven strengths that make active neurons quiescent at the next time (*t*: neuron* 1*;* $t+1$*: neurons* 4 *and* 5*;* $t+2$*: neurons* 2 *and* 3*). All of the gray arrows exhibit a fundamental law governing the interactions between nodal and coupling dynamics.*

and

$$US(t) < \sum_{j \in [x(t)]^-} b_j \quad \text{for each } t \in \Lambda^-.$$

Therefore

$$FS(t_*) + FS(t_*+1) + \cdots + FS(t^*-1) \geq \sum_{t \in \Lambda^+} \sum_{j \in [x(t)]^+} b_j \tag{8}$$

and

$$US(t_*) + US(t_*+1) + \cdots + US(t^*-1) < \sum_{t \in \Lambda^-} \sum_{j \in [x(t)]^-} b_j. \tag{9}$$

Since

$$\begin{aligned}&\sum_{t\in\Lambda^+}\sum_{j\in[x(t)]^+} b_j - \sum_{t\in\Lambda^-}\sum_{j\in[x(t)]^-} b_j\\ &= \sum_{t\in\Lambda^+}\sum_{j\in\mathbf{1}(x(t+1))\setminus\mathbf{1}(x(t))} b_j - \sum_{t\in\Lambda^-}\sum_{j\in\mathbf{1}(x(t))\setminus\mathbf{1}(x(t+1))} b_j\\ &= \sum_{t_*\le t<t^*}\langle x(t+1)-x(t), b\rangle = 0,\end{aligned}$$

inequality (5) follows from (8) and (9), and the proof is complete. □

5. Synaptic plasticity and neural synchrony. Theorem 1 exhibits a global, universal feature of the driving forces zeroing in on the order shared in the rise of diversity of network evolution. By contrast, the coincidence-detection evolving algorithm displays a local, uncertain feature of a generalized learning rule focusing on the time- and activity-dependent changes in coupling strengths. Combining the two distinct concepts gives a concise picture of regulation, concentrating on the assembling coordination of excitability within groups of neurons.

For this, define $E_U(t_*, t^*)$ to be $\sum_{i\in U}\min(\{a_{ii}(t);\ t = t_*, \ldots, t^*\})$, a quantity that measures the minimal total excitability within the group of neurons U in the period of time t_* and t^* with $t_* \le t^*$, where the coupling strength variable a_{ii} is considered to be a measure of excitability with respect to neuron i and, according to the working of neuron i, the increased excitability has a tendency to decrease the threshold for generating action potentials. Several lines of evidence in neuroscience have shown that activity-dependent modulation in intrinsic neuronal excitability could have a crucial role in modifying the integrative properties of neurons and their circuit dynamics [6, 10, 12, 26, 32, 51, 59]. This motivates us to use all those quantities $E_U(t_*, t^*)$ as an index for determining the existence of excitability coordination between groups of neurons, and we say that the minimal total excitability in the period of time $t = t_*, t_*+1, \ldots, t^*$ satisfies the *assembling coordination* if

$$E_U(t_*, t^*) \ge \sum_{i,j\in U}\max(\{a_{ij}(t) - a_{ji}(t);\ t = t_*, \ldots, t^*\}\cup\{0\}) \tag{10}$$

for each nonempty subset U of $\{1, 2, \ldots, n\}$.

Armed with the concept of assembling coordination, a solution to the synchronization problem, mentioned in section 2, may be stated as follows.

THEOREM 2. *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Given any initial neuronal active state $x(0)$ in the phase space $\{0,1\}^n$ and letting the discrete flow $x(t)$ iterate asynchronously, then a finite $T \ge 0$ can be determined so that if the minimal total excitability in the period of time $t = 0, 1, \ldots, T$ satisfies the assembling coordination, then a subset V of $\{1, 2, \ldots, n\}$ can be sorted out such that $\mathbf{1}(x(t)) = V$ for all $t \ge T$.*

Proof. Let $x(0)$ be any initial neuronal active state in $\{0,1\}^n$, and let $x(t)$ iterate asynchronously, guided by the dynamics (3), (4) and the coincidence-detection evolving algorithm.

We shall establish the following.

ASSERTION. *Given any $t_*, t^* = 0, 1, \ldots$ with $t_* \le t^*$, if the minimal total excitability in the period of time $t = t_*, t_*+1, \ldots, t^*$ fulfills the assembling coordination,* *then a feedback loop cannot occur in the fragment* $x(t_*), x(t_*+1), \ldots, x(t^*)$ *of the discrete flow* $x(t)$.

We will prove the assertion by arguing indirectly: assume that there exists $\hat{t}$ with $t_* < \hat{t} < t^*$ such that $x(t_*) = x(t^*) \neq x(\hat{t})$. The asynchronous updating of $x(t)$ implies that for every $t = t_*, t_* + 1, \ldots, t^* - 1$,

$$\sharp[x(t)]^+ + \sharp[x(t)]^- \leq 1. \tag{11}$$

Since

$$\sum_{t_* \leq t < t^*} (x_j(t+1) - x_j(t)) = 0 \tag{12}$$

for every $j = 1, 2, \ldots, n$, we deduce that

$$\bigcup_{t_* \leq t < t^*} [x(t)]^+ = \bigcup_{t_* \leq t < t^*} [x(t)]^-. \tag{13}$$

We can write (13) as a set of distinct elements

$$\{m_1, m_2, \ldots, m_q\}, \tag{14}$$

and for any $j = 1, 2, \ldots, q$ we put

$$M_j^+ = \{t;\ [x(t)]^+ = \{m_j\},\ t_* \leq t < t^*\},$$

$$M_j^- = \{t;\ [x(t)]^- = \{m_j\},\ t_* \leq t < t^*\}.$$

Then

$$\sharp M_j^+ = \sharp M_j^- \quad \text{for } j = 1, 2, \ldots, q \tag{15}$$

by (12). Consider the backward shift of the discrete flow $x(t)$ in the period of time t_* and t^*, and put

$$y(t) = x(t-1) \ \text{ for every } t = t_* + 1, \ldots, t^* \text{ and } y(t_*) = y(t^*).$$

For any $t = t_*, \ldots, t^* - 1$ we introduce two new quantities as follows:

$$\widetilde{FS}(t) = l_t(\mathbf{1}(y(t+1)), \mathbf{0}(y(t+1)) \cap \mathbf{1}(y(t)))$$

and

$$\widetilde{US}(t) = l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1)) \cap \mathbf{0}(y(t))).$$

A computation shows that

$$
\begin{aligned}
&\sum_{t_* \le t < t^*} (FS(t) - US(t)) + \sum_{t_* \le t < t^*} (\widetilde{FS}(t) - \widetilde{US}(t)) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^-) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1))) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^+) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^-) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^-) \qquad (16) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1))) \\
&= \sum_{t_* \le t < t^*} l_t([x(t)]^-, [x(t)]^+) + \sum_{t_* \le t < t^*} l_t([x(t)]^+, [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t([x(t)]^+, [x(t)]^+) - \sum_{t_* \le t < t^*} l_t([x(t)]^-, [x(t)]^-) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^-) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1)))
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{t_* \le t < t^*} (FS(t) - US(t)) - \sum_{t_* \le t < t^*} (\widetilde{FS}(t) - \widetilde{US}(t)) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1))) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), \mathbf{1}(x(t))) \qquad (17) \\
&\quad + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t([x(t)]^+, \mathbf{1}(x(t))) + \sum_{t_* \le t < t^*} l_t([x(t)]^-, \mathbf{1}(x(t))) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t))) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))).
\end{aligned}
$$

Claim 1.

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t([x(t)]^+,[x(t)]^+)+\sum_{t_*\le t<t^*} l_t([x(t)]^-,[x(t)]^-)\\ \ge&\sum_{1\le j,k\le q} 2\min(\{\sharp M_j^-,\sharp M_k^-\})\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\ldots,t^*\}\cup\{0\}).\end{aligned}$$

Proof of Claim 1. We first construct an array of dots as follows:

$$\begin{array}{ccccccc}
\bullet & \bullet & \bullet & \bullet & & \bullet & \bullet\\
\bullet & \bullet & & \bullet & & \bullet & \bullet\\
 & \bullet & & \bullet & \cdots & \bullet & \bullet\\
 & \bullet & & \bullet & & & \bullet\\
 & \bullet & & & & & \bullet\\
 & & & & & & \bullet
\end{array}$$

The array consists of q columns, q being the number of m_j's in (14). For each $j=1,2,\ldots,q$ the dots in the jth column are arranged consecutively from the top, and the number of dots in the jth column is equal to $\sharp M_j^-$, and so to $\sharp M_j^+$ by (15). Let r denote the number of rows in the array and for every $i=1,2,\ldots,r$ we define

$$V_i=\{j;\ \text{the } (i,j)\text{-entry of the array is equipped with } \bullet\ ,\ j=1,2,\ldots,q\}.$$

For each $i=1,2,\ldots,r$ and for each choice of $j\in V_i$, we assign to the dot in the ith row and jth column of the array, a pair (t_{ij},u_{ij}) with

$$t_{ij}\in M_j^+ \quad\text{and}\quad u_{ij}\in M_j^-,$$

such that

$$t_{1j}<t_{2j}<\cdots<t_{(\sharp M_j^+)j}$$

and

$$u_{1j}<u_{2j}<\cdots<u_{(\sharp M_j^-)j}.$$

So according to this construction and applying the assembling coordination (10) to $U=V_i$, we deduce that

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t([x(t)]^+,[x(t)]^+)+\sum_{t_*\le t<t^*} l_t([x(t)]^-,[x(t)]^-)\\
=&\sum_{1\le i\le r}\sum_{j\in V_i} a_{m_jm_j}(t_{ij})+\sum_{1\le i\le r}\sum_{j\in V_i} a_{m_jm_j}(u_{ij})\\
\ge&\sum_{1\le i\le r}\sum_{j\in V_i}\min(\{a_{m_jm_j}(t);\ t=t_*,\ldots,t^*\})+\sum_{1\le i\le r}\sum_{j\in V_i}\min(\{a_{m_jm_j}(t);\ t=t_*,\ldots,t^*\})\\
\ge&\sum_{1\le i\le r}\sum_{j,k\in V_i} 2\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\ldots,t^*\}\cup\{0\})\\
=&\sum_{1\le j,k\le q} 2\min(\{\sharp M_j^-,\sharp M_k^-\})\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\ldots,t^*\}\cup\{0\}),\end{aligned}$$

establishing Claim 1.

Claim 2.

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t)),[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t)),[x(t)]^-)\\&-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t)))+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t)))\\ \le&\sum_{1\le j,k\le q}\sum_{1\le i\le \sharp M_j^+}(x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij})-a_{m_km_j}(t_{ij})))\\&-\sum_{1\le j,k\le q}\sum_{1\le i\le \sharp M_j^-}(x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij})-a_{m_km_j}(u_{ij}))).\end{aligned}$$

Proof of Claim 2. For any $k\in\{1,2,\ldots,n\}\setminus\{m_1,m_2,\ldots,m_q\}$, we have either

$$\mathbf{1}(x(t))\cap\{k\}=\emptyset \quad\text{for all}\quad t=t_*,t_*+1,\ldots,t^*-1$$

or

$$\mathbf{1}(x(t))\cap\{k\}=\{k\} \quad\text{for all}\quad t=t_*,t_*+1,\ldots,t^*-1.$$

Case 1. $\mathbf{1}(x(t))\cap\{k\}=\emptyset$ *for all* $t=t_*,t_*+1,\ldots,t^*-1$. Then

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^-)\\&-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t))\cap\{k\})+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t))\cap\{k\})=0.\end{aligned}$$

Case 2. $\mathbf{1}(x(t))\cap\{k\}=\{k\}$ *for all* $t=t_*,t_*+1,\ldots,t^*-1$. Then

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^-)\\&-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t))\cap\{k\})+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t))\cap\{k\})\\=&\sum_{1\le j\le q}\sum_{1\le i\le \sharp M_j^-}(a_{m_jk}(t_{ij})-a_{m_jk}(u_{ij}))-\sum_{1\le j\le q}\sum_{1\le i\le \sharp M_j^-}(a_{km_j}(t_{ij})-a_{km_j}(u_{ij})).\end{aligned}$$

Fix $1\le j\le q$. Then (12) implies

$$\sum_{t_*\le t<t^*}(x_{m_j}(t+1)-x_{m_j}(t))=0,$$

and therefore

$$t_{1j}<u_{1j}<t_{2j}<u_{2j}<\cdots<t_{(\sharp M_j^+)j}<u_{(\sharp M_j^-)j} \tag{18}$$

or

$$u_{1j}<t_{1j}<u_{2j}<t_{2j}<\cdots<u_{(\sharp M_j^-)j}<t_{(\sharp M_j^+)j}. \tag{19}$$

Inequality (18) implies that for any $i = 1, 2, \ldots, \sharp M_j^+$,

$$\begin{aligned} (x_{m_j}(t_{ij}), x_k(t_{ij})) &= (0, 1), \\ (x_{m_j}(t_{ij}+1), x_k(t_{ij}+1)) &= (1, 1), \\ &\vdots \\ (x_{m_j}(u_{ij}-1), x_k(u_{ij}-1)) &= (1, 1), \\ (x_{m_j}(u_{ij}), x_k(u_{ij})) &= (1, 1). \end{aligned}$$

Thus for every $t = t_{ij}, t_{ij}+1, \ldots, u_{ij}-1$, we have

$$\delta_{m_j k}(t+1) = 1 \qquad \text{and} \qquad \delta_{k m_j}(t+1) = 0,$$

and according to the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)} a_{m_j k} \geq \mathcal{D}_{x(t)\to x(t+1)} a_{k m_j} \geq 0.$$

Therefore

$$\begin{aligned} &\sum_{1\leq i\leq \sharp M_j^-} (a_{m_j k}(t_{ij}) - a_{m_j k}(u_{ij})) - \sum_{1\leq i\leq \sharp M_j^-} (a_{k m_j}(t_{ij}) - a_{k m_j}(u_{ij})) \\ (20) \qquad &= \sum_{1\leq i\leq \sharp M_j^-} (-\mathcal{D}_{x(t_{ij})\to x(t_{ij}+1)} a_{m_j k} - \cdots - \mathcal{D}_{x(u_{ij}-1)\to x(u_{ij})} a_{m_j k}) \\ &\quad - \sum_{1\leq i\leq \sharp M_j^-} (-\mathcal{D}_{x(t_{ij})\to x(t_{ij}+1)} a_{k m_j} - \cdots - \mathcal{D}_{x(u_{ij}-1)\to x(u_{ij})} a_{k m_j}) \leq 0. \end{aligned}$$

On the other hand, (19) implies that for any $i = 1, 2, \ldots, \sharp M_j^-$,

$$\begin{aligned} (x_{m_j}(u_{ij}), x_k(u_{ij})) &= (1, 1), \\ (x_{m_j}(u_{ij}+1), x_k(u_{ij}+1)) &= (0, 1), \\ &\vdots \\ (x_{m_j}(t_{ij}-1), x_k(t_{ij}-1)) &= (0, 1), \\ (x_{m_j}(t_{ij}), x_k(t_{ij})) &= (0, 1). \end{aligned}$$

Thus for every $t = u_{ij}, u_{ij}+1, \ldots, t_{ij}-1$, we have

$$\delta_{m_j k}(t+1) = 1 \qquad \text{and} \qquad \delta_{k m_j}(t+1) = 0,$$

and by the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)} a_{m_j k} \leq \mathcal{D}_{x(t)\to x(t+1)} a_{k m_j} \leq 0.$$

Therefore

$$\begin{aligned} &\sum_{1\leq i\leq \sharp M_j^-} (a_{m_j k}(t_{ij}) - a_{m_j k}(u_{ij})) - \sum_{1\leq i\leq \sharp M_j^-} (a_{k m_j}(t_{ij}) - a_{k m_j}(u_{ij})) \\ (21) \qquad &= \sum_{1\leq i\leq \sharp M_j^-} (\mathcal{D}_{x(u_{ij})\to x(u_{ij}+1)} a_{m_j k} + \cdots + \mathcal{D}_{x(t_{ij}-1)\to x(t_{ij})} a_{m_j k}) \\ &\quad - \sum_{1\leq i\leq \sharp M_j^-} (\mathcal{D}_{x(u_{ij})\to x(u_{ij}+1)} a_{k m_j} + \cdots + \mathcal{D}_{x(t_{ij}-1)\to x(t_{ij})} a_{k m_j}) \leq 0. \end{aligned}$$

Inequalities (20) and (21) imply

$$\sum_{1\le j\le q}\sum_{1\le i\le \sharp M_j^-}(a_{m_jk}(t_{ij})-a_{m_jk}(u_{ij}))-\sum_{1\le j\le q}\sum_{1\le i\le \sharp M_j^-}(a_{km_j}(t_{ij})-a_{km_j}(u_{ij}))\le 0.$$

Combining Cases 1 and 2 gives

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap\{k\},[x(t)]^-)\\&-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t))\cap\{k\})+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t))\cap\{k\})\le 0\end{aligned}$$

for every $k\in\{1,2,\dots,n\}\setminus\{m_1,m_2,\dots,m_q\}$. Put

$$M=\{m_1,m_2,\dots,m_q\}.$$

Then

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t)),[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t)),[x(t)]^-)\\&\quad-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t)))+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t)))\\&\le\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap M,[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t))\cap M,[x(t)]^-)\\&\quad-\sum_{t_*\le t<t^*} l_t([x(t)]^+,\mathbf{1}(x(t))\cap M)+\sum_{t_*\le t<t^*} l_t([x(t)]^-,\mathbf{1}(x(t))\cap M)\\&=\sum_{1\le j\le q}\sum_{1\le i\le\sharp M_j^+}\sum_{1\le k\le q}(x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij})-a_{m_km_j}(t_{ij})))\\&\quad-\sum_{1\le j\le q}\sum_{1\le i\le\sharp M_j^-}\sum_{1\le k\le q}(x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij})-a_{m_km_j}(u_{ij})))\\&=\sum_{1\le j,k\le q}\sum_{1\le i\le\sharp M_j^+}(x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij})-a_{m_km_j}(t_{ij})))\\&\quad-\sum_{1\le j,k\le q}\sum_{1\le i\le\sharp M_j^-}(x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij})-a_{m_km_j}(u_{ij}))).\end{aligned}$$

Claim 3.

$$\begin{aligned}&\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t+1)),[x(t)]^+)-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t+1)),[x(t)]^-)\\&-\sum_{t_*\le t<t^*} l_t(\mathbf{1}(y(t+1)),\mathbf{1}(y(t+1)))+\sum_{t_*\le t<t^*} l_t(\mathbf{1}(x(t+1)),\mathbf{1}(x(t)))\le 0.\end{aligned}$$

Proof of Claim 3. Using $y(t) = x(t-1)$ for $t = t_* + 1, \dots, t^*$, we can derive the following string of identities:

$$\begin{aligned}
&\sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^+) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), [x(t)]^-) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(y(t+1)), \mathbf{1}(y(t+1))) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))) \\
&\quad - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), \mathbf{1}(x(t))) + \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t))) \\
&= \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) - \sum_{t_* \le t < t^*} l_t(\mathbf{1}(x(t)), \mathbf{1}(x(t))) \\
&= \sum_{t_* \le t < t^*-1} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) + l_{t^*-1}(\mathbf{1}(x(t^*)), \mathbf{1}(x(t^*))) \\
&\quad - l_{t_*}(\mathbf{1}(x(t_*)), \mathbf{1}(x(t_*))) - \sum_{t_* \le t < t^*-1} l_{t+1}(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) \\
&= \sum_{t_* \le t < t^*-1} l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) - \sum_{t_* \le t < t^*-1} l_{t+1}(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) \\
&\quad + \sum_{t_* \le t < t^*-1} l_{t+1}(\mathbf{1}(x(t_*)), \mathbf{1}(x(t_*))) - \sum_{t_* \le t < t^*-1} l_t(\mathbf{1}(x(t_*)), \mathbf{1}(x(t_*))).
\end{aligned}$$

Note that for any $t = 0, 1, \dots,$ we have

$$\begin{aligned}
&l_{t+1}(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) - l_t(\mathbf{1}(x(t+1)), \mathbf{1}(x(t+1))) \\
&= \sum_{i,j \in \mathbf{1}(x(t+1))} (a_{ij}(t+1) - a_{ij}(t)) \\
&= \sum_{i,j \in \mathbf{1}(x(t+1))} \mathcal{D}_{x(t) \to x(t+1)} a_{ij}
\end{aligned} \tag{22}$$

and

$$\begin{aligned}
&l_{t+1}(\mathbf{1}(x(t_*)), \mathbf{1}(x(t_*))) - l_t(\mathbf{1}(x(t_*)), \mathbf{1}(x(t_*))) \\
&= \sum_{i,j \in \mathbf{1}(x(t_*))} (a_{ij}(t+1) - a_{ij}(t)) \\
&= \sum_{i,j \in \mathbf{1}(x(t_*))} \mathcal{D}_{x(t) \to x(t+1)} a_{ij}.
\end{aligned} \tag{23}$$

The coincidence-detection evolving algorithm implies that for every $t = 0, 1, \dots$ and $i, j = 1, 2, \dots, n$,

$$\text{if } \mathcal{D}_{x(t) \to x(t+1)} a_{ij} > 0, \text{ then } i, j \in \mathbf{1}(x(t+1)),$$

$$\text{if } \mathcal{D}_{x(t) \to x(t+1)} a_{ij} < 0, \text{ then either } i \notin \mathbf{1}(x(t+1)) \text{ or } j \notin \mathbf{1}(x(t+1)),$$

so that

$$\sum_{i,j \in \mathbf{1}(x(t+1))} \mathcal{D}_{x(t) \to x(t+1)} a_{ij} \ge \sum_{i,j \in \mathbf{1}(x(t_*))} \mathcal{D}_{x(t) \to x(t+1)} a_{ij}. \tag{24}$$

Combining (22), (23), and (24) gives

$$\begin{aligned}&\sum_{t_*\le t<t^*-1} l_t(\mathbf{1}(x(t+1)),\mathbf{1}(x(t+1))) - \sum_{t_*\le t<t^*-1} l_{t+1}(\mathbf{1}(x(t+1)),\mathbf{1}(x(t+1)))\\ &+\sum_{t_*\le t<t^*-1} l_{t+1}(\mathbf{1}(x(t_*)),\mathbf{1}(x(t_*))) - \sum_{t_*\le t<t^*-1} l_t(\mathbf{1}(x(t_*)),\mathbf{1}(x(t_*))) \le 0.\end{aligned}$$

Composing the a priori estimates in Claims 1, 2, and 3 with the inequality

$$\begin{aligned}&\sum_{1\le j,k\le q}\ \sum_{1\le i\le \sharp M_j^+} (x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij})-a_{m_km_j}(t_{ij})))\\ (25)\quad &-\sum_{1\le j,k\le q}\ \sum_{1\le i\le \sharp M_j^-} (x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij})-a_{m_km_j}(u_{ij})))\\ &\le \sum_{1\le j,k\le q} 2\min(\{\sharp M_j^-,\sharp M_k^-\})\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\dots,t^*\}\cup\{0\}),\end{aligned}$$

we conclude from (11), (16), and (17) that

$$\begin{aligned}&2((FS(t_*)+\cdots+FS(t^*-1))-(US(t_*)+\cdots+US(t^*-1)))\\ &=\sum_{t_*\le t<t^*}(FS(t)-US(t))+\sum_{t_*\le t<t^*}(\widetilde{FS}(t)-\widetilde{US}(t))\\ &\quad+\sum_{t_*\le t<t^*}(FS(t)-US(t))-\sum_{t_*\le t<t^*}(\widetilde{FS}(t)-\widetilde{US}(t))\\ &\le-\sum_{1\le j,k\le q} 2\min(\{\sharp M_j^-,\sharp M_k^-\})\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\dots,t^*\}\cup\{0\})\\ &\quad+\sum_{1\le j,k\le q}\ \sum_{1\le i\le \sharp M_j^+} (x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij})-a_{m_km_j}(t_{ij})))\\ &\quad-\sum_{1\le j,k\le q}\ \sum_{1\le i\le \sharp M_j^-} (x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij})-a_{m_km_j}(u_{ij})))\le 0,\end{aligned}$$

in contradiction to Theorem 1. So to complete the proof of the assertion, it remains to show that (25) holds. To see this, fix $1\le j,k\le q$ with $j\ne k$, and think of the set $M_k^+\cup M_k^-$ as the holes in a sieve that filters the set $M_j^+\cup M_j^-$. Thus we divide the sets

$$\begin{aligned}M_j^+&=\{t_{1j},t_{2j},\dots,t_{(\sharp M_j^+)j}\},\\ M_j^-&=\{u_{1j},u_{2j},\dots,u_{(\sharp M_j^-)j}\}\end{aligned}$$

into mutually disjoint ν classes

$$E_1^j,E_2^j,\dots,E_\nu^j$$

such that

(a) $E_1^j\cup E_2^j\cup\cdots\cup E_\nu^j=M_j^+\cup M_j^-$,

(b) $\max(E_\eta^j)<\min(E_{\eta+1}^j)$ for $\eta=1,2,\dots,\nu-1$,

(c) there does not exist t in $M_k^+\cup M_k^-$ such that $\min(E_\eta^j)\le t\le\max(E_\eta^j)$ for $\eta=1,2,\dots,\nu$,

(d) there exists t in $M_k^+ \cup M_k^-$ such that $\max(E_\eta^j) < t < \min(E_{\eta+1}^j)$ for $\eta = 1, 2, \ldots, \nu - 1$.

According to (a) we have

$$\begin{aligned}
&\sum_{1\le i\le \sharp M_j^+} (x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij}) - a_{m_km_j}(t_{ij}))) \\
&- \sum_{1\le i\le \sharp M_j^-} (x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij}) - a_{m_km_j}(u_{ij}))) \\
&= \sum_{1\le\eta\le\nu} \sum_{t\in M_j^+\cap E_\eta^j} (x_{m_k}(t)(a_{m_jm_k}(t) - a_{m_km_j}(t))) \\
&- \sum_{1\le\eta\le\nu} \sum_{t\in M_j^-\cap E_\eta^j} (x_{m_k}(t)(a_{m_jm_k}(t) - a_{m_km_j}(t))).
\end{aligned} \tag{26}$$

Fix $1 \le \eta \le \nu$. Then (c) implies that either

$$x_{m_k}(t) = 0 \quad \text{for } \min(E_\eta^j) \le t \le \max(E_\eta^j) \tag{27}$$

or

$$x_{m_k}(t) = 1 \quad \text{for } \min(E_\eta^j) \le t \le \max(E_\eta^j). \tag{28}$$

Case 1. *$\sharp E_\eta^j$ is even.* According to (27) and (28) we have to distinguish between two subcases.

Subcase 1-1. *$x_{m_k}(t) = 0$ for $\min(E_\eta^j) \le t \le \max(E_\eta^j)$.* Then

$$\sum_{t\in M_j^+\cap E_\eta^j} (x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))-\sum_{t\in M_j^-\cap E_\eta^j} (x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t))) = 0.$$

Subcase 1-2. *$x_{m_k}(t) = 1$ for $\min(E_\eta^j) \le t \le \max(E_\eta^j)$.* According to (b) we can write E_η^j as

$$\{t_1, t_2, \ldots, t_{\sharp E_\eta^j/2}, u_1, u_2, \ldots, u_{\sharp E_\eta^j/2}\},$$

where $t_i \in M_j^+$ and $u_i \in M_j^-$ for $i = 1, 2, \ldots, \sharp E_\eta^j/2$, such that either

$$t_1 < u_1 < t_2 < u_2 < \cdots < t_{\sharp E_\eta^j/2} < u_{\sharp E_\eta^j/2} \tag{29}$$

or

$$u_1 < t_1 < u_2 < t_2 < \cdots < u_{\sharp E_\eta^j/2} < t_{\sharp E_\eta^j/2}. \tag{30}$$

Inequality (29) implies that for any $i = 1, 2, \ldots, \sharp E_\eta^j/2$,

$$\begin{aligned}
(x_{m_j}(t_i), x_{m_k}(t_i)) &= (0, 1), \\
(x_{m_j}(t_i + 1), x_{m_k}(t_i + 1)) &= (1, 1), \\
&\vdots \\
(x_{m_j}(u_i - 1), x_{m_k}(u_i - 1)) &= (1, 1), \\
(x_{m_j}(u_i), x_{m_k}(u_i)) &= (1, 1).
\end{aligned}$$

Thus for every $t = t_i, t_i + 1, \dots, u_i - 1$, we have

$$\delta_{m_j m_k}(t+1) = 1 \qquad \text{and} \qquad \delta_{m_k m_j}(t+1) = 0,$$

and according to the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)} a_{m_j m_k} \geq \mathcal{D}_{x(t)\to x(t+1)} a_{m_k m_j} \geq 0.$$

Therefore

$$\begin{aligned}
&\sum_{t\in M_j^+ \cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t))) - \sum_{t\in M_j^- \cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t)))\\
&= \sum_{1\leq i\leq \sharp E_\eta^j/2} (a_{m_j m_k}(t_i) - a_{m_j m_k}(u_i)) - \sum_{1\leq i\leq \sharp E_\eta^j/2} (a_{m_k m_j}(t_i) - a_{m_k m_j}(u_i))\\
&= \sum_{1\leq i\leq \sharp E_\eta^j/2} (-\mathcal{D}_{x(t_i)\to x(t_i+1)} a_{m_j m_k} - \dots - \mathcal{D}_{x(u_i-1)\to x(u_i)} a_{m_j m_k})\\
&\quad - \sum_{1\leq i\leq \sharp E_\eta^j/2} (-\mathcal{D}_{x(t_i)\to x(t_i+1)} a_{m_k m_j} - \dots - \mathcal{D}_{x(u_i-1)\to x(u_i)} a_{m_k m_j}) \leq 0.
\end{aligned}$$

On the other hand, (30) implies that for any $i = 1, 2, \dots, \sharp E_\eta^j/2$,

$$\begin{aligned}
(x_{m_j}(u_i), x_{m_k}(u_i)) &= (1,1),\\
(x_{m_j}(u_i+1), x_{m_k}(u_i+1)) &= (0,1),\\
&\vdots\\
(x_{m_j}(t_i-1), x_{m_k}(t_i-1)) &= (0,1),\\
(x_{m_j}(t_i), x_{m_k}(t_i)) &= (0,1).
\end{aligned}$$

Thus for every $t = u_i, u_i + 1, \dots, t_i - 1$, we have

$$\delta_{m_j m_k}(t+1) = 1 \qquad \text{and} \qquad \delta_{m_k m_j}(t+1) = 0,$$

and according to the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)} a_{m_j m_k} \leq \mathcal{D}_{x(t)\to x(t+1)} a_{m_k m_j} \leq 0.$$

Therefore

$$\begin{aligned}
&\sum_{t\in M_j^+ \cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t))) - \sum_{t\in M_j^- \cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t)))\\
&= \sum_{1\leq i\leq \sharp E_\eta^j/2} (a_{m_j m_k}(t_i) - a_{m_j m_k}(u_i)) - \sum_{1\leq i\leq \sharp E_\eta^j/2} (a_{m_k m_j}(t_i) - a_{m_k m_j}(u_i))\\
&= \sum_{1\leq i\leq \sharp E_\eta^j/2} (\mathcal{D}_{x(u_i)\to x(u_i+1)} a_{m_j m_k} + \dots + \mathcal{D}_{x(t_i-1)\to x(t_i)} a_{m_j m_k})\\
&\quad - \sum_{1\leq i\leq \sharp E_\eta^j/2} (\mathcal{D}_{x(u_i)\to x(u_i+1)} a_{m_k m_j} + \dots + \mathcal{D}_{x(t_i-1)\to x(t_i)} a_{m_k m_j}) \leq 0.
\end{aligned}$$

Case 2. $\sharp E^j_\eta$ *is odd.*

Subcase 2-1. $x_{m_k}(t) = 0$ *for* $\min(E^j_\eta) \le t \le \max(E^j_\eta)$. Then

$$\sum_{t\in M^+_j\cap E^j_\eta}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t))) - \sum_{t\in M^-_j\cap E^j_\eta}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t))) = 0.$$

Subcase 2-2. $x_{m_k}(t) = 1$ *for* $\min(E^j_\eta) \le t \le \max(E^j_\eta)$. According to (b) we can write E^j_η as

$$\{t_1, t_2, \ldots, t_{(\sharp E^j_\eta-1)/2}, t_{(\sharp E^j_\eta+1)/2}, u_1, u_2, \ldots, u_{(\sharp E^j_\eta-1)/2}\},$$

where $t_i \in M^+_j$ for $i = 1, 2, \ldots, (\sharp E^j_\eta+1)/2$, $u_i \in M^-_j$ for $i = 1, 2, \ldots, (\sharp E^j_\eta-1)/2$ and

$$t_1 < u_1 < t_2 < u_2 < \cdots < t_{(\sharp E^j_\eta-1)/2} < u_{(\sharp E^j_\eta-1)/2} < t_{(\sharp E^j_\eta+1)/2}, \tag{31}$$

or as

$$\{t_1, t_2, \ldots, t_{(\sharp E^j_\eta-1)/2}, u_1, u_2, \ldots, u_{(\sharp E^j_\eta-1)/2}, u_{(\sharp E^j_\eta+1)/2}\},$$

where $t_i \in M^+_j$ for $i = 1, 2, \ldots, (\sharp E^j_\eta-1)/2$, $u_i \in M^-_j$ for $i = 1, 2, \ldots, (\sharp E^j_\eta+1)/2$ and

$$u_1 < t_1 < u_2 < t_2 < \cdots < u_{(\sharp E^j_\eta-1)/2} < t_{(\sharp E^j_\eta-1)/2} < u_{(\sharp E^j_\eta+1)/2}. \tag{32}$$

Inequality (31) implies that for any $i = 1, 2, \ldots, (\sharp E^j_\eta-1)/2$,

$$\begin{aligned}(x_{m_j}(t_i), x_{m_k}(t_i)) &= (0, 1),\\ (x_{m_j}(t_i+1), x_{m_k}(t_i+1)) &= (1, 1),\\ &\vdots\\ (x_{m_j}(u_i-1), x_{m_k}(u_i-1)) &= (1, 1),\\ (x_{m_j}(u_i), x_{m_k}(u_i)) &= (1, 1).\end{aligned}$$

Thus for every $t = t_i, t_i+1, \ldots, u_i-1$, we have

$$\delta_{m_jm_k}(t+1) = 1 \qquad \text{and} \qquad \delta_{m_km_j}(t+1) = 0,$$

and by the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)}a_{m_jm_k} \ge \mathcal{D}_{x(t)\to x(t+1)}a_{m_km_j} \ge 0.$$

Then

$$\begin{aligned}&\sum_{t\in M^+_j\cap E^j_\eta}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t))) - \sum_{t\in M^-_j\cap E^j_\eta}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\\ &= \sum_{1\le i\le(\sharp E^j_\eta-1)/2}(a_{m_jm_k}(t_i)-a_{m_jm_k}(u_i)) - \sum_{1\le i\le(\sharp E^j_\eta-1)/2}(a_{m_km_j}(t_i)-a_{m_km_j}(u_i))\\ &\quad + a_{m_jm_k}(t_{(\sharp E^j_\eta+1)/2}) - a_{m_km_j}(t_{(\sharp E^j_\eta+1)/2})\\ &= \sum_{1\le i\le(\sharp E^j_\eta-1)/2}(-\mathcal{D}_{x(t_i)\to x(t_i+1)}a_{m_jm_k} - \cdots - \mathcal{D}_{x(u_i-1)\to x(u_i)}a_{m_jm_k})\end{aligned}$$

$$
\begin{aligned}
&- \sum_{1\le i\le (\sharp E_\eta^j-1)/2} (-\mathcal{D}_{x(t_i)\to x(t_i+1)} a_{m_k m_j} - \cdots - \mathcal{D}_{x(u_i-1)\to x(u_i)} a_{m_k m_j})\\
&+ a_{m_j m_k}(t_{(\sharp E_\eta^j+1)/2}) - a_{m_k m_j}(t_{(\sharp E_\eta^j+1)/2})\\
\le\ & a_{m_j m_k}(t_{(\sharp E_\eta^j+1)/2}) - a_{m_k m_j}(t_{(\sharp E_\eta^j+1)/2})\\
\le\ & \max(\{a_{m_j m_k}(t) - a_{m_k m_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}).
\end{aligned}
$$

On the other hand, (32) implies that for any $i = 1, 2, \dots, (\sharp E_\eta^j - 1)/2$,

$$
\begin{aligned}
(x_{m_j}(u_i), x_{m_k}(u_i)) &= (1,1),\\
(x_{m_j}(u_i+1), x_{m_k}(u_i+1)) &= (0,1),\\
&\vdots\\
(x_{m_j}(t_i-1), x_{m_k}(t_i-1)) &= (0,1),\\
(x_{m_j}(t_i), x_{m_k}(t_i)) &= (0,1).
\end{aligned}
$$

Thus for every $t = u_i, u_i+1, \dots, t_i-1$, we have

$$\delta_{m_j m_k}(t+1) = 1 \qquad \text{and} \qquad \delta_{m_k m_j}(t+1) = 0,$$

and using the coincidence-detection evolving algorithm, we conclude that

$$\mathcal{D}_{x(t)\to x(t+1)} a_{m_j m_k} \le \mathcal{D}_{x(t)\to x(t+1)} a_{m_k m_j} \le 0.$$

Then

$$
\begin{aligned}
&\sum_{t\in M_j^+\cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t))) - \sum_{t\in M_j^-\cap E_\eta^j} (x_{m_k}(t)(a_{m_j m_k}(t) - a_{m_k m_j}(t)))\\
=\ & \sum_{1\le i\le (\sharp E_\eta^j-1)/2} (a_{m_j m_k}(t_i) - a_{m_j m_k}(u_i)) - \sum_{1\le i\le (\sharp E_\eta^j-1)/2} (a_{m_k m_j}(t_i) - a_{m_k m_j}(u_i))\\
&- a_{m_j m_k}(u_{(\sharp E_\eta^j+1)/2}) + a_{m_k m_j}(u_{(\sharp E_\eta^j+1)/2})\\
=\ & \sum_{1\le i\le (\sharp E_\eta^j-1)/2} (\mathcal{D}_{x(u_i)\to x(u_i+1)} a_{m_j m_k} + \cdots + \mathcal{D}_{x(t_i-1)\to x(t_i)} a_{m_j m_k})\\
&- \sum_{1\le i\le (\sharp E_\eta^j-1)/2} (\mathcal{D}_{x(u_i)\to x(u_i+1)} a_{m_k m_j} + \cdots + \mathcal{D}_{x(t_i-1)\to x(t_i)} a_{m_k m_j})\\
&- a_{m_j m_k}(u_{(\sharp E_\eta^j+1)/2}) + a_{m_k m_j}(u_{(\sharp E_\eta^j+1)/2})\\
\le\ & a_{m_k m_j}(u_{(\sharp E_\eta^j+1)/2}) - a_{m_j m_k}(u_{(\sharp E_\eta^j+1)/2})\\
\le\ & \max(\{a_{m_k m_j}(t) - a_{m_j m_k}(t);\ t = t_*, \dots, t^*\} \cup \{0\}).
\end{aligned}
$$

Let

$$\Delta = \{\eta;\ \sharp E_\eta^j \text{ is odd},\ \eta = 1, 2, \dots, \nu\}.$$

Combining Cases 1 and 2 gives

$$\begin{aligned}
&\sum_{1\le\eta\le\nu}\sum_{t\in M_j^+\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\\
&\quad-\sum_{1\le\eta\le\nu}\sum_{t\in M_j^-\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\\
&=\sum_{\eta\notin\Delta}\left(\sum_{t\in M_j^+\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\right.\\
&\qquad\left.-\sum_{t\in M_j^-\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\right)\\
&\quad+\sum_{\eta\in\Delta}\left(\sum_{t\in M_j^+\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\right.\\
&\qquad\left.-\sum_{t\in M_j^-\cap E_\eta^j}(x_{m_k}(t)(a_{m_jm_k}(t)-a_{m_km_j}(t)))\right)\\
&\le\sum_{\eta\in\Delta,\max(E_\eta^j)\in M_j^+}\max(\{a_{m_jm_k}(t)-a_{m_km_j}(t);\ t=t_*,\dots,t^*\}\cup\{0\})\\
&\quad+\sum_{\eta\in\Delta,\max(E_\eta^j)\in M_j^-}\max(\{a_{m_km_j}(t)-a_{m_jm_k}(t);\ t=t_*,\dots,t^*\}\cup\{0\}).
\end{aligned}\tag{33}$$

Since

$$\sharp E_1^j+\sharp E_2^j+\cdots+\sharp E_\nu^j=\sharp(M_j^+\cup M_j^-),$$

$\sharp\Delta$ is even. Let us write

$$\Delta=\{\eta_1,\eta_2,\dots,\eta_{2\beta}\},$$

where $\beta\ge 0$ and $\eta_1<\eta_2<\cdots<\eta_{2\beta}$. Then

$$2\beta\le\nu\le 2\sharp M_j^-$$

and (d) gives

$$\nu-1\le\sharp(M_k^+\cup M_k^-).$$

Hence

$$\beta\le\frac{1}{2}\nu\le\min\left(\left\{\sharp M_j^-,\frac{1}{2}+\sharp M_k^-\right\}\right),$$

and so

$$\beta\le\min(\{\sharp M_j^-,\sharp M_k^-\}).$$

Note that for each $i = 1, 2, \dots, 2\beta - 1$, if $\max(E^j_{\eta_i})$ belongs to M^+_j (resp., M^-_j), then $\max(E^j_{\eta_{i+1}})$ belongs to M^-_j (resp., M^+_j). Thus

$$\begin{aligned}
&\sum_{\eta\in\Delta,\max(E^j_\eta)\in M^+_j} \max(\{a_{m_jm_k}(t) - a_{m_km_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&\quad + \sum_{\eta\in\Delta,\max(E^j_\eta)\in M^-_j} \max(\{a_{m_km_j}(t) - a_{m_jm_k}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&= \beta \max(\{a_{m_jm_k}(t) - a_{m_km_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \qquad (34)\\
&\quad + \beta \max(\{a_{m_km_j}(t) - a_{m_jm_k}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&\le \min(\{\sharp M^-_j, \sharp M^-_k\}) \max(\{a_{m_jm_k}(t) - a_{m_km_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&\quad + \min(\{\sharp M^-_j, \sharp M^-_k\}) \max(\{a_{m_km_j}(t) - a_{m_jm_k}(t);\ t = t_*, \dots, t^*\} \cup \{0\}).
\end{aligned}$$

So we conclude from (26), (33), and (34) that

$$\begin{aligned}
&\sum_{1\le j,k\le q} \sum_{1\le i\le \sharp M^+_j} (x_{m_k}(t_{ij})(a_{m_jm_k}(t_{ij}) - a_{m_km_j}(t_{ij}))) \\
&\quad - \sum_{1\le j,k\le q} \sum_{1\le i\le \sharp M^-_j} (x_{m_k}(u_{ij})(a_{m_jm_k}(u_{ij}) - a_{m_km_j}(u_{ij}))) \\
&\le \sum_{1\le j,k\le q} \min(\{\sharp M^-_j, \sharp M^-_k\}) \max(\{a_{m_jm_k}(t) - a_{m_km_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&\quad + \sum_{1\le j,k\le q} \min(\{\sharp M^-_j, \sharp M^-_k\}) \max(\{a_{m_km_j}(t) - a_{m_jm_k}(t);\ t = t_*, \dots, t^*\} \cup \{0\}) \\
&= \sum_{1\le j,k\le q} 2\min(\{\sharp M^-_j, \sharp M^-_k\}) \max(\{a_{m_jm_k}(t) - a_{m_km_j}(t);\ t = t_*, \dots, t^*\} \cup \{0\}),
\end{aligned}$$

establishing (25) and completing the proof of the assertion.

Since the discrete flow $x(t)$ iterates asynchronously, there exists a sequence of time steps $T_1, T_2, \dots$ with $0 < T_1 < T_2 < \cdots$ such that

$$\bigcup_{T_j\le t<T_{j+1}} s(t) = \{1, 2, \dots, n\} \tag{35}$$

for any $j = 1, 2, \dots$. The assertion and the above choice of $T_1, T_2, \dots$ in effect imply that a finite $k > 1$ and a subset V of $\{1, 2, \dots, n\}$ can be determined so that if the minimal total excitability in the period of time $t = 0, 1, \dots, T_k$ satisfies the assembling coordination, then

$$\mathbf{1}(x(t)) = V \quad \text{for all} \quad T_{k-1} \le t \le T_k. \tag{36}$$

(A passing remark: Without the coincidence-detection evolving algorithm support, (3), (4), and (36) cannot generically arrive at the conclusion that $\mathbf{1}(x(t)) = V$ for all $t \ge T_k$.) Set

$$T' = T_{k-1} \quad \text{and} \quad T = T_k.$$

Armed with (36), we now claim that for all $t \geq T$ the discrete flow $x(t)$ undergoing the coincidence-detection evolving algorithm satisfies

$$\mathbf{1}(x(t)) = V. \tag{37}$$

For the proof, put

$$\gamma(t) = \left(\mathbf{1}(x(t)) \cap \left\{ i; \sum_j a_{ij}(t)x_j(t) < b_i \right\}\right) \cup \left(\mathbf{0}(x(t)) \cap \left\{ i; \sum_j a_{ij}(t)x_j(t) \geq b_i \right\}\right)$$

for $t = 0, 1, \ldots$. Then, by (36) and the coincidence-detection evolving algorithm, we have for any given $i = 1, 2, \ldots, n$ and $T' < t \leq T$

$$\begin{aligned} &\sum_{1\leq j\leq n} a_{ij}(t)x_j(t) \\ &= \sum_{1\leq j\leq n} (a_{ij}(t-1) + \mathcal{D}_{x(t-1)\to x(t)} a_{ij})x_j(t) \\ &= \sum_{1\leq j\leq n} a_{ij}(t-1)x_j(t-1) + \sum_{1\leq j\leq n} \mathcal{D}_{x(t-1)\to x(t)} a_{ij} x_j(t) \\ &\geq \sum_{1\leq j\leq n} a_{ij}(t-1)x_j(t-1) \end{aligned}$$

if $i \in \mathbf{1}(x(t))$; otherwise

$$\begin{aligned} &\sum_{1\leq j\leq n} a_{ij}(t)x_j(t) \\ &= \sum_{1\leq j\leq n} (a_{ij}(t-1) + \mathcal{D}_{x(t-1)\to x(t)} a_{ij})x_j(t) \\ &= \sum_{1\leq j\leq n} a_{ij}(t-1)x_j(t-1) + \sum_{1\leq j\leq n} \mathcal{D}_{x(t-1)\to x(t)} a_{ij} x_j(t) \\ &\leq \sum_{1\leq j\leq n} a_{ij}(t-1)x_j(t-1). \end{aligned}$$

Thus we have

$$\gamma(T') \supset \gamma(T'+1) \supset \cdots \supset \gamma(T). \tag{38}$$

Furthermore, $\gamma(T) = \emptyset$. To see this, suppose $\gamma(T) \neq \emptyset$. Then (35) and (38) imply that there is τ with $T' \leq \tau < T$ such that $\gamma(\tau) \cap s(\tau) \neq \emptyset$, and so $\mathbf{1}(x(\tau+1)) \neq V$, which contradicts (36). We apply now the condition $\gamma(T) = \emptyset$ to prove the assertion (37), which shows that a transient period of synchronization of neural firing *propagates* to the whole period of time $t \geq T$. (As illustrated in Figure 4, the condition $\gamma(T) = \emptyset$ means that a transition state of neural activity occurs at time T. The accumulation of both excitability coordination and activity-dependent changes of coupling strengths causes a group of neurons to come into synchronized activity prior to time T. When a transient period of synchronization of neural firing occurs, we need only the support of coincidence detection to produce a sort of positive feedback that admits the synchronized neural impulses between populations of neurons to continue to synchronize posterior to time T.)

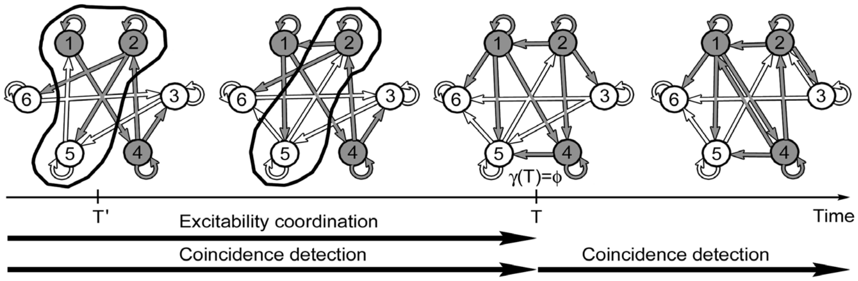


FIG. 4. *A monotonically decreasing sequence of $\gamma(t)$ (the nodes encompassed by closed curves in the period of time T' and T) can be generated in a transient period of synchronization of neural firing. The condition $\gamma(T) = \emptyset$ determines a transition state of neural activity at time T.*

We prove the assertion (37) by induction on time $t \geq T$. The case of $t = T$ follows from (36). Assume that $t > T$ and the assertion (37) is true for all cases of time less than t (except the cases of time less than T). Then, by induction hypothesis and the coincidence-detection evolving algorithm, we have

$$\gamma(T) \supset \gamma(T+1) \supset \cdots \supset \gamma(t-1).$$

Since $\gamma(T) = \emptyset$, we get $\gamma(t-1) = \emptyset$, so that

$$\sum_j a_{ij}(t-1)x_j(t-1) \geq b_i \quad \text{for all} \quad i \in \mathbf{1}(x(t-1)), \tag{39}$$

$$\sum_j a_{ij}(t-1)x_j(t-1) < b_i \quad \text{for all} \quad i \in \mathbf{0}(x(t-1)). \tag{40}$$

Inequalities (39), (40) and the induction hypothesis together imply that $\mathbf{1}(x(t)) = V$. This completes the inductive proof of the assertion (37) and concludes the proof of the theorem. □

6. Stability of neural synchrony. This section is devoted to the study of the stability problem of neural synchrony underlying the nonlinear dynamical system modeled by the parametric equations (3) and (4). The question may be stated as follows: *Would small perturbations of the initial neuronal active state cause only small variations of the discrete flow which iterates to a state of synchronous neuronal firing?*

To solve this, we first introduce a quantity to clarify the mathematical meaning of disturbance of neuronal activity states, and then introduce the phenomenon of local absorption of the discrete flow.

For any given n-by-n real matrix $A = [a_{ij}]_{n\times n}$ and $s \in \{1, 2, \ldots, n\}$, the *state transition function* $H_{A,s} : \{0,1\}^n \longrightarrow \{0,1\}^n$ is defined by $[H_{A,s}(x)]_i = hea(\sum_{j=1}^n a_{ij}x_j - b_i)$ if $i = s$; otherwise x_i, $i = 1, 2, \ldots, n$. For each pair x, y of distinct points of $\{0,1\}^n$, we define the *proximal number from x to y* with respect to the state transition functions $H_{A,\cdot}$ by

$$Pro_A(x,y) = \min\{r;\ H_{A,s_{r-1}} \circ H_{A,s_{r-2}} \circ \cdots \circ H_{A,s_0}(x) = y\},$$

and let $Pro_A(x,x) = 0$. Here the operation " ∘ " denotes the composition of two functions. For each nonempty subset Ω of $\{0,1\}^n$, the *proximal number from x to Ω*

with respect to the state transition functions $H_{A,\boldsymbol{\cdot}}$ is defined to be the minimum of proximal numbers $Pro_A(x, y)$, where y is taken over all elements in Ω. The proximal number $Pro_A(x, y)$ measures the absorption of x into y underlying the state transition functions $H_{A,\boldsymbol{\cdot}}$.

Let $x(t)$ iterate according to the parametric equations (3) and (4). We show now that, based on the coincidence-detection evolving algorithm, the establishment of the evolutionary couplings enables the states locally absorbed by $x(t)$ to be absorbed by $x(t+1)$ for all $t = 0, 1, \ldots$. This phenomenon is aptly called *local absorption* of the discrete flow $x(t)$, which reveals an underlying principle of the coincidence-detection rule to stabilize neural synchrony.

Theorem 3. *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Let the discrete flow $x(t)$ iterate asynchronously and let $A(t)$ satisfy the condition of assembling coordination described in Theorem* 2. *If the plasticity parameters satisfy*

$$\left| \sum_{j \in \mathbf{1}(x(t+1)), j \neq i} \mathcal{D}_{x(t) \to x(t+1)} a_{ij} \right| \geq \max_{j \in \mathbf{0}(x(t+1))} |\mathcal{D}_{x(t) \to x(t+1)} a_{ij}| \tag{41}$$

for all $i \in \mathbf{1}(x(t+1))$ and $t = 0, 1, \ldots,$ then

$$\{y;\ Pro_{A(t)}(y, x(t)) \leq 1\} \subset \{y;\ Pro_{A(t+1)}(y, x(t+1)) \leq 2\} \tag{42}$$

for all $t = 0, 1, \ldots$.

Proof. Let $x(0)$ be any initial neuronal active state in $\{0, 1\}^n$, and let $x(t)$ iterate asynchronously, guided by the dynamics (3), (4) and the coincidence-detection evolving algorithm. Fix $\tau \geq 0$ and consider the fragment $x(\tau), x(\tau+1), x(\tau+2), x(\tau+3)$ of the discrete flow $x(t)$. We may first suppose that, in the period of time $\tau, \tau+1, \tau+2$, the minimal total excitability fulfills the assembling coordination and the plasticity parameters satisfy (41). According to (3) and (4), there are $A(\tau), A(\tau+1), A(\tau+2)$ and $s(\tau), s(\tau+1), s(\tau+2)$ such that

$$x(t+1) = H_{A(t), s(t)}(x(t)) \quad \text{for } t = \tau, \tau+1, \tau+2,$$

and by (4) we have

$$l_{t+1}(\mathbf{1}(x(\tau)), s(\tau)) = l_t(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j \in \mathbf{1}(x(\tau))} \mathcal{D}_{x(t) \to x(t+1)} a_{s(\tau)j} \tag{43}$$

for $t = \tau, \tau+1$. To prove (42), we have to claim that

$$x(\tau+1) = H_{A(\tau+1), s(\tau)}(x(\tau)) \tag{44}$$

and

$$x(\tau+1) = H_{A(\tau+2), s(\tau)}(x(\tau)). \tag{45}$$

To prove (44), we consider (43) at time $t = \tau$, that is,

$$l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) = l_\tau(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j \in \mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau) \to x(\tau+1)} a_{s(\tau)j}.$$

By the asynchronous iteration of $x(t)$, we split the arguments into three cases.

Case 1. $\mathbf{1}(x(\tau)) \subsetneq \mathbf{1}(x(\tau+1))$. Then $s(\tau) = \mathbf{0}(x(\tau)) \cap \mathbf{1}(x(\tau+1))$, and according to the coincidence-detection evolving algorithm, we have

$$\mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \geq 0 \quad \text{for all } j \in \mathbf{1}(x(\tau)). \tag{46}$$

Since $x(\tau+1) = H_{A(\tau),s(\tau)}(x(\tau))$, we have

$$l_\tau(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}, \tag{47}$$

and hence, together with (46), we conclude that

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) &= l_\tau(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \\ &\geq l_\tau(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}. \end{aligned} \tag{48}$$

This implies that

$$H_{A(\tau+1),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (44).

Case 2. $\mathbf{1}(x(\tau)) \supsetneq \mathbf{1}(x(\tau+1))$. Then $s(\tau) = \mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+1))$, and according to the coincidence-detection evolving algorithm, we have

$$\mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \leq 0 \quad \text{for all } j \in \mathbf{1}(x(\tau)). \tag{49}$$

Since $x(\tau+1) = H_{A(\tau),s(\tau)}(x(\tau))$, we have

$$l_\tau(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}, \tag{50}$$

and hence, together with (49), we conclude that

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) &= l_\tau(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \\ &\leq l_\tau(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}. \end{aligned} \tag{51}$$

This implies that

$$H_{A(\tau+1),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (44).

Case 3. $\mathbf{1}(x(\tau)) = \mathbf{1}(x(\tau+1))$. Then either $s(\tau) \in \mathbf{1}(x(\tau))$ or $s(\tau) \in \mathbf{0}(x(\tau))$. If $s(\tau) \in \mathbf{1}(x(\tau))$, then, according to the coincidence-detection evolving algorithm, we have

$$\mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \geq 0 \quad \text{for all } j \in \mathbf{1}(x(\tau)). \tag{52}$$

Since $x(\tau+1) = H_{A(\tau),s(\tau)}(x(\tau))$, we have

$$l_\tau(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}, \tag{53}$$

and hence, together with (52), we conclude that

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) &= l_\tau(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \\ &\geq l_\tau(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}. \end{aligned} \tag{54}$$

This implies that

$$H_{A(\tau+1),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1).$$

On the other hand, if $s(\tau) \in \mathbf{0}(x(\tau))$, then, according to the coincidence-detection evolving algorithm, we have

$$\mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \le 0 \quad \text{for all } j \in \mathbf{1}(x(\tau)). \tag{55}$$

Since $x(\tau+1) = H_{A(\tau),s(\tau)}(x(\tau))$, we have

$$l_\tau(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}, \tag{56}$$

and hence, together with (55), we conclude that

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) &= l_\tau(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s(\tau)j} \\ &\le l_\tau(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}. \end{aligned} \tag{57}$$

This implies that

$$H_{A(\tau+1),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (44).

We turn now to establish (45). Since $x(t)$ iterates asynchronously, we need only consider three cases.

Case 1. $\mathbf{1}(x(\tau)) \subsetneq \mathbf{1}(x(\tau+1))$. Then

$$s(\tau) = \mathbf{0}(x(\tau)) \cap \mathbf{1}(x(\tau+1)) \tag{58}$$

and, by (43) and (48), we have

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &\ge b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}. \end{aligned} \tag{59}$$

Subcase 1-1. $\mathbf{1}(x(\tau+1)) \subsetneq \mathbf{1}(x(\tau+2))$. Then, by (58), we have

$$s(\tau) \in \mathbf{1}(x(\tau+2)) \quad \text{and} \quad \mathbf{1}(x(\tau)) \subset \mathbf{1}(x(\tau+2)),$$

and according to the coincidence-detection evolving algorithm, we see that

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \ge 0 \quad \text{for all } j \in \mathbf{1}(x(\tau)).$$

Thus, by (59), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) \ge b_{s(\tau)}.$$

Combining this with (47) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 1-2. $\mathbf{1}(x(\tau+1)) \supsetneq \mathbf{1}(x(\tau+2))$. Then $s(\tau) \neq s(\tau+1)$. Indeed, if $s(\tau) = s(\tau+1)$, then, applying (48), (58), and the assembling coordination (10) to $U = s(\tau)$, $t_* = \tau$, and $t^* = \tau+2$, we get

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau+1)) &= l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau)) \\ &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + l_{\tau+1}(s(\tau), s(\tau)) \\ &\geq b_{s(\tau)} + a_{s(\tau)s(\tau)}(\tau+1) \\ &\geq b_{s(\tau)}. \end{aligned} \tag{60}$$

Since $s(\tau+1) = s(\tau) \in \mathbf{1}(x(\tau+1))$, inequality (60) implies that

$$x(\tau+2) = H_{A(\tau+1), s(\tau+1)}(x(\tau+1)) = x(\tau+1),$$

contradicting the assumption $\mathbf{1}(x(\tau+1)) \supsetneq \mathbf{1}(x(\tau+2))$. Now combining $s(\tau) \neq s(\tau+1)$ with the fact that

$$s(\tau) \in \mathbf{1}(x(\tau+1)) \quad \text{and} \quad s(\tau+1) = \mathbf{1}(x(\tau+1)) \cap \mathbf{0}(x(\tau+2))$$

gives

$$s(\tau) \in \mathbf{1}(x(\tau+2)). \tag{61}$$

Thus, together with (58), we have

$$\begin{aligned} s(\tau+1) &= \mathbf{1}(x(\tau+1)) \cap \mathbf{0}(x(\tau+2)) \\ &= (\mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+2))) \cup (s(\tau) \cap \mathbf{0}(x(\tau+2))) \\ &= \mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+2)) \end{aligned} \tag{62}$$

and

$$\begin{aligned} \mathbf{1}(x(\tau+2)) \setminus s(\tau) &= (\mathbf{1}(x(\tau+1)) \cap \mathbf{1}(x(\tau+2))) \setminus (s(\tau) \cap \mathbf{1}(x(\tau+2))) \\ &= (\mathbf{1}(x(\tau+1)) \setminus s(\tau)) \cap \mathbf{1}(x(\tau+2)) \\ &= \mathbf{1}(x(\tau)) \cap \mathbf{1}(x(\tau+2)). \end{aligned} \tag{63}$$

From (62) and (63), we see that

$$\begin{aligned} &\sum_{j \in \mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)j} \\ &= \sum_{j \in \mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+2))} \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)j} + \sum_{j \in \mathbf{1}(x(\tau)) \cap \mathbf{1}(x(\tau+2))} \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)j} \\ &= \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)s(\tau+1)} + \sum_{j \in \mathbf{1}(x(\tau+2)), j \neq s(\tau)} \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)j}, \end{aligned}$$

and hence, by (41), (61), and the coincidence-detection evolving algorithm, we have

$$\sum_{j \in \mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1) \to x(\tau+2)} a_{s(\tau)j} \geq 0.$$

Thus, by (59), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}.$$

Combining this with (47) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 1-3. $\mathbf{1}(x(\tau+1)) = \mathbf{1}(x(\tau+2))$. Then, by (58), we have

$$s(\tau) \in \mathbf{1}(x(\tau+2)) \quad \text{and} \quad \mathbf{1}(x(\tau)) \subset \mathbf{1}(x(\tau+2)),$$

and according to the coincidence-detection evolving algorithm, we see that

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \geq 0 \ \text{ for all } j \in \mathbf{1}(x(\tau)).$$

Thus, by (59), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}.$$

Combining this with (47) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Case 2. $\mathbf{1}(x(\tau)) \supsetneq \mathbf{1}(x(\tau+1))$. Then

$$s(\tau) = \mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+1)) \tag{64}$$

and, by (43) and (51), we have

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &< b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}. \end{aligned} \tag{65}$$

Subcase 2-1. $\mathbf{1}(x(\tau+1)) \subsetneq \mathbf{1}(x(\tau+2))$. Then $s(\tau) \neq s(\tau+1)$. Indeed, if $s(\tau) = s(\tau+1)$, then, applying (51), (64), and the assembling coordination (10) to $U = s(\tau)$, $t_* = \tau$, and $t^* = \tau+2$, we get

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau+1)) &= l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau)) \\ &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) - l_{\tau+1}(s(\tau), s(\tau)) \\ &< b_{s(\tau)} - a_{s(\tau)s(\tau)}(\tau+1) \\ &\leq b_{s(\tau)}. \end{aligned} \tag{66}$$

Since $s(\tau+1) = s(\tau) \in \mathbf{0}(x(\tau+1))$, inequality (66) implies that

$$x(\tau+2) = H_{A(\tau+1),s(\tau+1)}(x(\tau+1)) = x(\tau+1),$$

contradicting the assumption $\mathbf{1}(x(\tau+1)) \supsetneq \mathbf{1}(x(\tau+2))$. Now combining $s(\tau) \neq s(\tau+1)$ with the fact that

$$s(\tau) \in \mathbf{0}(x(\tau+1)) \quad \text{and} \quad s(\tau+1) = \mathbf{0}(x(\tau+1)) \cap \mathbf{1}(x(\tau+2))$$

gives

$$s(\tau) \in \mathbf{0}(x(\tau+2)). \tag{67}$$

From (67) and the coincidence-detection evolving algorithm, we see that

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \leq 0 \ \text{ for all } j \in \mathbf{1}(x(\tau)).$$

Thus, by (65), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}.$$

Combining this with (50) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 2-2. $\mathbf{1}(x(\tau+1)) \supsetneq \mathbf{1}(x(\tau+2))$. Then, by (64), we have $s(\tau) \in \mathbf{0}(x(\tau+2))$, and hence, according to the coincidence-detection evolving algorithm, we see that

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \leq 0 \ \text{ for all } j \in \mathbf{1}(x(\tau)).$$

Thus, by (65), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}.$$

Combining this with (50) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 2-3. $\mathbf{1}(x(\tau+1)) = \mathbf{1}(x(\tau+2))$. Then, by (64), we have $s(\tau) \in \mathbf{0}(x(\tau+2))$, and hence, according to the coincidence-detection evolving algorithm, we see that

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \leq 0 \ \text{ for all } j \in \mathbf{1}(x(\tau)).$$

Thus, by (65), we have

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) < b_{s(\tau)}.$$

Combining this with (50) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Case 3. $\mathbf{1}(x(\tau)) = \mathbf{1}(x(\tau+1))$. Then the assertion holds: *If* $x(\tau+1) \neq x(\tau+2)$, *then* $s(\tau) \neq s(\tau+1)$. Indeed, if $s(\tau) = s(\tau+1)$, then, by (54) and (57), we have

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau+1)) &= l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau)) \\ &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) \\ &\geq b_{s(\tau)} \quad \text{if } s(\tau) \in \mathbf{1}(x(\tau)) \end{aligned} \tag{68}$$

and

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau+1)) &= l_{\tau+1}(\mathbf{1}(x(\tau+1)), s(\tau)) \\ &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) \\ &< b_{s(\tau)} \quad \text{if } s(\tau) \in \mathbf{0}(x(\tau)). \end{aligned} \tag{69}$$

Inequalities (68) and (69) together imply that

$$x(\tau+2)=H_{A(\tau+1),s(\tau+1)}(x(\tau+1))=x(\tau+1),$$

contradicting the assumption $x(\tau{+}1)\neq x(\tau{+}2)$. This contradiction shows the validity of the assertion, and next we have to consider three subcases.

Subcase 3-1. $\mathbf{1}(x(\tau+1))\subsetneq \mathbf{1}(x(\tau+2))$. Then

$$s(\tau+1)=\mathbf{0}(x(\tau+1))\cap\mathbf{1}(x(\tau+2)) \tag{70}$$

and, according to the assertion, we have

$$s(\tau)\neq s(\tau+1). \tag{71}$$

Since $\mathbf{1}(x(\tau))=\mathbf{1}(x(\tau+1))$, we have either

$$s(\tau)\in\mathbf{1}(x(\tau))=\mathbf{1}(x(\tau+1)) \tag{72}$$

or

$$s(\tau)\in\mathbf{0}(x(\tau))=\mathbf{0}(x(\tau+1)). \tag{73}$$

In case of (72), we see that

$$s(\tau)\in\mathbf{1}(x(\tau+2)) \quad\text{and}\quad \mathbf{1}(x(\tau))\subset\mathbf{1}(x(\tau+2)), \tag{74}$$

and from (43) and (54) we have

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)),s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)),s(\tau))+\sum_{j\in\mathbf{1}(x(\tau))}\mathcal{D}_{x(\tau+1)\to x(\tau+2)}a_{s(\tau)j}\\ &\geq b_{s(\tau)}+\sum_{j\in\mathbf{1}(x(\tau))}\mathcal{D}_{x(\tau+1)\to x(\tau+2)}a_{s(\tau)j}. \end{aligned} \tag{75}$$

Thus we conclude from (74), (75), and the coincidence-detection evolving algorithm that

$$l_{\tau+2}(\mathbf{1}(x(\tau)),s(\tau))\geq b_{s(\tau)}.$$

Combining this with (53) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau))=H_{A(\tau),s(\tau)}(x(\tau))=x(\tau+1).$$

On the other hand, in case of (73) we have

$$s(\tau)\in\mathbf{0}(x(\tau+2))$$

by combining (70), (71), and (73). Hence, by the coincidence-detection evolving algorithm, we have

$$\mathcal{D}_{x(\tau+1)\to x(\tau+2)}a_{s(\tau)j}\leq 0 \quad\text{for all } j\in\mathbf{1}(x(\tau)). \tag{76}$$

Combining (76) with (43) and (57) implies that

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)),s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)),s(\tau))+\sum_{j\in\mathbf{1}(x(\tau))}\mathcal{D}_{x(\tau+1)\to x(\tau+2)}a_{s(\tau)j}\\ &< b_{s(\tau)}+\sum_{j\in\mathbf{1}(x(\tau))}\mathcal{D}_{x(\tau+1)\to x(\tau+2)}a_{s(\tau)j}\\ &\leq b_{s(\tau)}. \end{aligned}$$

Thus, together with (56), we have

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 3-2. $\mathbf{1}(x(\tau+1)) \supsetneq \mathbf{1}(x(\tau+2))$. Then

$$s(\tau+1) = \mathbf{1}(x(\tau+1)) \cap \mathbf{0}(x(\tau+2)) \tag{77}$$

and, according to the assertion, we have

$$s(\tau) \neq s(\tau+1). \tag{78}$$

Since $\mathbf{1}(x(\tau)) = \mathbf{1}(x(\tau+1))$, we have either

$$s(\tau) \in \mathbf{1}(x(\tau)) = \mathbf{1}(x(\tau+1)) \tag{79}$$

or

$$s(\tau) \in \mathbf{0}(x(\tau)) = \mathbf{0}(x(\tau+1)). \tag{80}$$

In case of (79), we have

$$s(\tau) \in \mathbf{1}(x(\tau+2)) \tag{81}$$

by combining (77), (78), and (79). Since

$$\begin{aligned}\mathbf{1}(x(\tau)) \cap \mathbf{1}(x(\tau+2)) &= \mathbf{1}(x(\tau+1)) \cap \mathbf{1}(x(\tau+2))\\ &= \mathbf{1}(x(\tau+2))\end{aligned}$$

and

$$\begin{aligned}\mathbf{1}(x(\tau)) \cap \mathbf{0}(x(\tau+2)) &= \mathbf{1}(x(\tau+1)) \cap \mathbf{0}(x(\tau+2))\\ &= s(\tau+1),\end{aligned}$$

we conclude from (43) and (54) that

$$\begin{aligned}l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}\\ &\geq b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau))\cap\mathbf{1}(x(\tau+2))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}\\ &\quad + \sum_{j\in\mathbf{1}(x(\tau))\cap\mathbf{0}(x(\tau+2))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}\\ &= b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau+2))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j}\\ &\quad + \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)s(\tau+1)}.\end{aligned} \tag{82}$$

According to (41), (81), and the coincidence-detection evolving algorithm, we get

$$\begin{aligned}&\sum_{j\in\mathbf{1}(x(\tau+2))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} + \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)s(\tau+1)}\\ &= \sum_{j\in\mathbf{1}(x(\tau+2)),j\neq s(\tau)} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} + \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)s(\tau)}\\ &\quad + \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)s(\tau+1)}\\ &\geq \sum_{j\in\mathbf{1}(x(\tau+2)),j\neq s(\tau)} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} + \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)s(\tau+1)} \geq 0.\end{aligned} \tag{83}$$

Thus, from (82) and (83), we see that

$$l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) \geq b_{s(\tau)}.$$

Combining this with (53) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1).$$

On the other hand, in case of (80) we have

$$s(\tau) \in \mathbf{0}(x(\tau+2))$$

and, from (43), (57), and the coincidence-detection evolving algorithm, we see that

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &< b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &\leq b_{s(\tau)}. \end{aligned}$$

Combining this with (56) implies that

$$H_{A(\tau+2),s(\tau)}(x(\tau)) = H_{A(\tau),s(\tau)}(x(\tau)) = x(\tau+1),$$

proving (45).

Subcase 3-3. $\mathbf{1}(x(\tau+1)) = \mathbf{1}(x(\tau+2))$. Then, by (43), (54), and the coincidence-detection evolving algorithm, we have

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &\geq b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau+2))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &\geq b_{s(\tau)} \quad \text{if } s(\tau) \in \mathbf{1}(x(\tau)), \end{aligned} \tag{84}$$

and, on the other hand, by (43), (57), and the coincidence-detection evolving algorithm we have

$$\begin{aligned} l_{\tau+2}(\mathbf{1}(x(\tau)), s(\tau)) &= l_{\tau+1}(\mathbf{1}(x(\tau)), s(\tau)) + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &< b_{s(\tau)} + \sum_{j\in\mathbf{1}(x(\tau))} \mathcal{D}_{x(\tau+1)\to x(\tau+2)} a_{s(\tau)j} \\ &\leq b_{s(\tau)} \quad \text{if } s(\tau) \in \mathbf{0}(x(\tau)) = \mathbf{0}(x(\tau+2)). \end{aligned} \tag{85}$$

Combining (84) and (85) accordingly with (53) and (56) implies that

$$x(\tau+2) = H_{A(\tau+1),s(\tau+1)}(x(\tau+1)) = x(\tau+1),$$

proving (45).

Having completed the proof of the claims (44) and (45), we turn now to establish (42). To see this, let $T \geq 0$ be determined by Theorem 2 so that the minimal total excitability in the period of time $t = 0, 1, \ldots, T$ satisfies the assembling coordination and that

$$\mathbf{1}(x(t)) = V \quad \text{for all} \quad t \geq T.$$

Given $\tau = 0, 1, \ldots, T-1$ and $y \in \{y;\ Pro_{A(\tau)}(y, x(\tau)) \leq 1\}$, then either

$$y = x(\tau) \tag{86}$$

or

$$H_{A(\tau),(\mathbf{1}(x(\tau))\cap\mathbf{0}(y))\cup(\mathbf{0}(x(\tau))\cap\mathbf{1}(y))}(y) = x(\tau). \tag{87}$$

In case of (86), it is readily seen that

$$y \in \{y;\ Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2\}$$

since $H_{A(\tau+1),s(\tau)}(x(\tau)) = x(\tau+1)$ by (44). In case of (87), we consider the discrete flow $z(t)$ given by $z(0) = y$, $z(1) = x(\tau)$, $z(2) = x(\tau+1), \ldots$ so that

$$H_{W(t),\widetilde{s}(t)}(z(t)) = z(t+1) \quad \text{for } t = 0, 1, \ldots,$$

where

$$W(0) = A(\tau), W(1) = A(\tau), W(2) = A(\tau+1), \ldots$$

and

$$\widetilde{s}(0) = (\mathbf{1}(x(\tau)) \cap \mathbf{0}(y)) \cup (\mathbf{0}(x(\tau)) \cap \mathbf{1}(y)), \widetilde{s}(1) = s(\tau), \widetilde{s}(2) = s(\tau+1), \ldots.$$

Let $\mathcal{D}_{w(t)\to w(t+1)} w_{ij} = w_{ij}(t+1) - w_{ij}(t)$ for all $i, j \in \{1, 2, \ldots, n\}$ and $t = 0, 1, \ldots$. Then, based on the construction of $x(t)$, $A(t)$, and $W(0) = W(1)$, it is readily seen that $z(t)$ will be one of the discrete flows guided also by the dynamics (3), (4) and the coincidence-detection evolving algorithm. Further, since $\tau + 1 \leq T$ and the assembling coordination, associated to $x(t)$, is satisfied in the period of time $t = 0, 1, \ldots, T$, we have

$$\begin{aligned}\sum_{i\in U} \min(\{w_{ii}(t);\ t = 0, 1, 2\}) &\geq \sum_{i\in U} \min(\{a_{ii}(t);\ t = 0, \ldots, T\}) \\ &\geq \sum_{i,j\in U} \max(\{a_{ij}(t) - a_{ji}(t);\ t = 0, \ldots, T\} \cup \{0\}) \\ &\geq \sum_{i,j\in U} \max(\{w_{ij}(t) - w_{ji}(t);\ t = 0, 1, 2\} \cup \{0\})\end{aligned}$$

for each nonempty subset U of $\{1, 2, \ldots, n\}$. Thus, according to (45) and the fact that the choice of $\mathcal{D}_{w(t)\to w(t+1)} w_{ij}$ satisfies (41), we have

$$H_{W(2),\widetilde{s}(0)}(z(0)) = z(1). \tag{88}$$

Since

$$H_{W(2),\widetilde{s}(1)}(z(1)) = z(2) \tag{89}$$

by (44), we conclude from (88) and (89) that

$$H_{A(\tau+1),\widetilde{s}(1)} \circ H_{A(\tau+1),\widetilde{s}(0)}(y) = x(\tau+1),$$

proving (42) for $t = 0, 1, \ldots, T-1$. Now let $\tau \geq T$ be given and consider y satisfying $Pro_{A(\tau)}(y, x(\tau)) \leq 1$. If $y = x(\tau)$, then

$$Pro_{A(\tau+1)}(y, x(\tau+1)) = Pro_{A(\tau+1)}(x(\tau), x(\tau)) = 0,$$

so

$$y \in \{y;\ Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2\}.$$

If $y \neq x(\tau)$, then $Pro_{A(\tau)}(y, x(\tau)) = 1$, and hence exactly one of the following holds:

$$\mathbf{1}(x(\tau)) \cap \mathbf{0}(y) \neq \emptyset \quad \text{or} \quad \mathbf{0}(x(\tau)) \cap \mathbf{1}(y) \neq \emptyset. \tag{90}$$

Since $Pro_{A(\tau)}(y, x(\tau)) = 1$, the former of (90) implies that

$$l_\tau(\mathbf{1}(y), \mathbf{1}(x(\tau)) \cap \mathbf{0}(y)) \geq b_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)}.$$

Thus

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(y), \mathbf{1}(x(\tau)) \cap \mathbf{0}(y)) &= l_\tau(\mathbf{1}(y), \mathbf{1}(x(\tau)) \cap \mathbf{0}(y)) \\ &\quad + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)j} \\ &\geq b_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)} + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)j}. \end{aligned}$$

Since $\mathbf{0}(x(\tau)) \cap \mathbf{1}(y) = \emptyset$ and $\mathbf{1}(x(\tau)) = \mathbf{1}(x(\tau+1))$, we have

$$\mathcal{D}_{x(\tau)\to x(\tau+1)} a_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)j} \geq 0 \quad \text{for all } j \in \mathbf{1}(y).$$

This implies that

$$l_{\tau+1}(\mathbf{1}(y), \mathbf{1}(x(\tau)) \cap \mathbf{0}(y)) \geq b_{\mathbf{1}(x(\tau))\cap\mathbf{0}(y)},$$

and hence

$$H_{A(\tau+1),\mathbf{1}(x(\tau))\cap\mathbf{0}(y)}(y) = x(\tau) = x(\tau+1).$$

So we have

$$Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2.$$

On the other hand, suppose the latter of (90) holds. Then

$$l_\tau(\mathbf{1}(y), \mathbf{0}(x(\tau)) \cap \mathbf{1}(y)) < b_{\mathbf{0}(x(\tau))\cap\mathbf{1}(y)}.$$

Since $\mathbf{0}(x(\tau)) = \mathbf{0}(x(\tau+1))$, we have

$$\begin{aligned} l_{\tau+1}(\mathbf{1}(y), \mathbf{0}(x(\tau)) \cap \mathbf{1}(y)) &= l_\tau(\mathbf{1}(y), \mathbf{0}(x(\tau)) \cap \mathbf{1}(y)) \\ &\quad + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{\mathbf{0}(x(\tau))\cap\mathbf{1}(y)j} \\ &< b_{\mathbf{0}(x(\tau))\cap\mathbf{1}(y)} + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{\mathbf{0}(x(\tau))\cap\mathbf{1}(y)j} \\ &\leq b_{\mathbf{0}(x(\tau))\cap\mathbf{1}(y)}. \end{aligned}$$

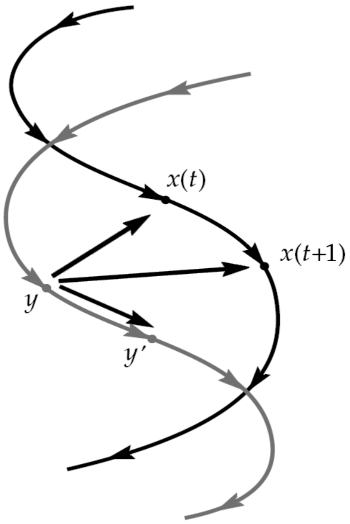


FIG. 5. *A schematic illustration of what local absorption effects. Consider the discrete flow $x(t)$ (the black curve) and its perturbation (the gray curve). Based on the local absorption of the discrete flow $x(t)$, each state y absorbed by $x(t)$ underlying $H_{A(t),\cdot}$ (the arrow from y to $x(t)$) can also be absorbed by $x(t+1)$ underlying $H_{A(t+1),\cdot}$ (the arrow from y to $x(t+1)$). Thus, as the state y transits (the arrow from y to y'), it can only cause small variations of the state $x(t+1)$.*

This implies that

$$H_{A(\tau+1),\mathbf{0}(x(\tau))\cap\mathbf{1}(y)}(y) = x(\tau) = x(\tau+1),$$

completing the proof of (42) for $t = T, T+1, \ldots,$ and the proof of Theorem 3 is complete. □

Theorem 3 is applied to show that neural synchrony is stable (see Figure 5 for an illustration). To accomplish this, let $x(t)$, $A(t)$, and $s(t)$ be defined by (3) and (4), and consider the perturbed system

$$(91) \qquad y(t+1) = H_{A(t),\widetilde{s}(t)}(y(t)), \quad t = 0, 1, \ldots,$$

where $\widetilde{s}(t) \in \{1, 2, \ldots, n\}$ for $t = 0, 1, \ldots.$

THEOREM 4. *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Let the discrete flow $x(t)$ iterate asynchronously, and let $A(t)$ satisfy the condition of assembling coordination described in Theorem* 2. *If the regime of plasticity parameters* (41) *holds, then for any $y(0) \in \{0,1\}^n$ with*

$$(92) \qquad Pro_{A(0)}(y(0), x(0)) \leq 2,$$

there are discrete flows $y(t)$ for the perturbed system (91) *such that*

$$Pro_{A(t)}(y(t), x(t)) \leq 2$$

for all $t = 0, 1, \ldots.$

Proof. Choose $y(0) \in \{0,1\}^n$ so that

$$Pro_{A(0)}(y(0), x(0)) \leq 2.$$

Having chosen $y(0), y(1), \ldots, y(\tau)$, it is readily seen that if $Pro_{A(\tau)}(y(\tau), x(\tau)) = 0$, then

$$\{H_{A(\tau),s}(y(\tau)), s = 1, 2, \ldots, n\} \cap \{y;\ Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2\} \neq \emptyset;$$

and if $Pro_{A(\tau)}(y(\tau), x(\tau)) \neq 0$, then, by (42), we have

$$\begin{aligned} &\{H_{A(\tau),s}(y(\tau)), s = 1, 2, \ldots, n\} \cap \{y;\ Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2\} \\ \supset\ &\{H_{A(\tau),s}(y(\tau)), s = 1, 2, \ldots, n\} \cap \{y;\ Pro_{A(\tau)}(y, x(\tau)) \leq 1\} \neq \emptyset. \end{aligned}$$

Choose $y(\tau+1)$ in

$$\{H_{A(\tau),s}(y(\tau)), s = 1, 2, \ldots, n\} \cap \{y;\ Pro_{A(\tau+1)}(y, x(\tau+1)) \leq 2\}.$$

Then

$$Pro_{A(\tau+1)}(y(\tau+1), x(\tau+1)) \leq 2,$$

and further we have

$$y(\tau+1) \in \{H_{A(\tau),s}(y(\tau)), s = 1, 2, \ldots, n\}.$$

Thus the discrete flow $y(t)$ is constructed for the perturbed system (91) such that

$$Pro_{A(t)}(y(t), x(t)) \leq 2 \quad \text{for all } t = 0, 1, \ldots,$$

and the proof is complete. $\square$

7. Nonlinear effect of neural synchrony. Synchrony and stability of synchrony may lead to formulating evolutionary network architecture. To visualize this, we first show that the effect of synchronization admits self-sustaining activity of strengthening in evolutionary couplings, and then show that such strengthening gives the robust stability of neural synchrony.

The first of these follows immediately from the coincidence-detection evolving algorithm. In fact, we have shown in Theorem 2 that the discrete flow $x(t)$ can iterate to a state x^* of synchronous neuronal firing, that is, a finite $T \geq 0$ can be determined so that

$$x(t) = x^* \quad \text{for all } t \geq T. \tag{93}$$

By the alternating nature of (3) and (4), and by the coincidence-detection evolving algorithm, we see that the assertion (93) is equivalent to saying that for each $t \geq T$, we have the following chain of implications:

$$\begin{aligned} & x(t) = x(t+1) = x^* \\ \Longrightarrow & \begin{cases} a_{ij}(t+1) - a_{ij}(t) = \mathcal{D}_{x(t) \to x(t+1)} a_{ij} \geq 0 & \text{if } i, j \in \mathbf{1}(x(t+1)) = \mathbf{1}(x^*), \\ a_{ij}(t+1) - a_{ij}(t) = \mathcal{D}_{x(t) \to x(t+1)} a_{ij} \leq 0 & \text{otherwise.} \end{cases} \\ \Longrightarrow & x(t+1) = x(t+2) = x^* \\ \Longrightarrow & \begin{cases} a_{ij}(t+2) - a_{ij}(t+1) = \mathcal{D}_{x(t+1) \to x(t+2)} a_{ij} \geq 0 & \text{if } i, j \in \mathbf{1}(x(t+2)) = \mathbf{1}(x^*), \\ a_{ij}(t+2) - a_{ij}(t+1) = \mathcal{D}_{x(t+1) \to x(t+2)} a_{ij} \leq 0 & \text{otherwise.} \end{cases} \\ \Longrightarrow & x(t+2) = x(t+3) = x^* \\ \Longrightarrow & \cdots. \end{aligned}$$

The chain of implications demonstrates how synchrony could indeed specify positive feedback. It is positive feedback to give rise to the consolidation of sync-dependent circuitry, which feeds back to reinforce the neurons to fire in synchrony.

To see the second of these, let us consider the evolutionary coupling states $A(T)$, $A(T+1), \ldots$, each containing distinctive sync-dependent circuitry resulting from the positive feedback. Given a fixed $A(\tau),\ \tau = T, T+1, \ldots$, it follows from (39) and (40) that for any choice of $s \in \{1, 2, \ldots, n\}$,

$$H_{A(\tau),s}(x^*) = x^*.$$

Thus x^* is a *common fixed point* of state transition functions $H_{A(\tau),s}$ for all $s = 1, 2, \ldots, n$, and x^* corresponds to an equilibrium state of the dynamical system being modeled by the nonlinear parametric equations

$$\begin{aligned} z(t+1) &= H_{W(t),s(t)}(z(t)), \quad t = 0, 1, \ldots, \\ W(t+1) &= W(t) + D_{z(t)\to z(t+1)}W, \quad t = 0, 1, \ldots, \end{aligned} \tag{94}$$

where $W(0) = A(\tau)$ and $z(t)$ iterates asynchronously, guided by the coincidence-detection evolving algorithm. By Theorem 4, we can associate to each $A(\tau)$ a region

$$\{y;\ Pro_{A(\tau)}(y, x^*) \le 2\} \tag{95}$$

such that for any $y(0)$ chosen from (95), there are discrete flows $y(t)$ for the perturbed system of (94) satisfying

$$Pro_{W(t)}(y(t), x^*) \le 2 \ \text{ for all } t = 0, 1, \ldots. \tag{96}$$

Therefore, (96) indicates that to every distinctive construction of sync-dependent circuitry $A(\tau)$, $\tau = T,\ T+1,\ \ldots$, there corresponds the region of states (95) which initializes the stability process of x^*.

With the notion above, we can show that, under the conditions of Theorem 4, the inclusions

$$\{y;\ Pro_{A(T)}(y, x^*) \le 2\} \subset \{y;\ Pro_{A(T+1)}(y, x^*) \le 2\} \subset \cdots \tag{97}$$

hold. This reveals the *robust stability* of neural synchrony, meaning that the synchronization state x^* is not only stable but also capable of expanding the region of states for initializing its stability process. For the proof of (97), let $\tau \ge T$ be given and y satisfy $Pro_{A(\tau)}(y, x^*) \le 2$. If $Pro_{A(\tau)}(y, x^*) \le 1$, then, by Theorem 3, we have

$$Pro_{A(\tau+1)}(y, x^*) \le 2.$$

On the other hand, if $Pro_{A(\tau)}(y, x^*) = 2$, then there exist $s_0, s_1 \in \{1, 2, \ldots, n\}$ with $s_0 \ne s_1$ such that

$$H_{A(\tau),s_1} \circ H_{A(\tau),s_0}(y) = x^*. \tag{98}$$

These y, $H_{A(\tau),s_0}(y)$, and x^* are mutually distinct. Let $y' = H_{A(\tau),s_0}(y)$. We have to show that

$$y' = H_{A(\tau+1),s_0}(y) \quad \text{and} \quad x^* = H_{A(\tau+1),s_1}(y'). \tag{99}$$

Case 1. $s_0, s_1 \in \mathbf{1}(x^*)$. Then, by (98), we have

$$\mathbf{1}(y) \subsetneq \mathbf{1}(y') \subsetneq \mathbf{1}(x^*),$$

and

$$s_0 = \mathbf{0}(y) \cap \mathbf{1}(y') \quad \text{and} \quad s_1 = \mathbf{0}(y') \cap \mathbf{1}(x^*).$$

This implies that

$$l_\tau(\mathbf{1}(y), s_0) \geq b_{s_0} \quad \text{and} \quad l_\tau(\mathbf{1}(y'), s_1) \geq b_{s_1}.$$

Hence, according to the coincidence-detection evolving algorithm, we have

$$l_{\tau+1}(\mathbf{1}(y), s_0) = l_\tau(\mathbf{1}(y), s_0) + \sum_{j \in \mathbf{1}(y)} \mathcal{D}_{x(\tau) \to x(\tau+1)} a_{s_0 j} \geq b_{s_0}$$

and

$$l_{\tau+1}(\mathbf{1}(y'), s_1) = l_\tau(\mathbf{1}(y'), s_1) + \sum_{j \in \mathbf{1}(y')} \mathcal{D}_{x(\tau) \to x(\tau+1)} a_{s_1 j} \geq b_{s_1},$$

proving (99).

Case 2. $s_0, s_1 \in \mathbf{0}(x^*)$. Then, by (98), we have

$$\mathbf{0}(y) \subsetneq \mathbf{0}(y') \subsetneq \mathbf{0}(x^*),$$

and

$$s_0 = \mathbf{1}(y) \cap \mathbf{0}(y') \quad \text{and} \quad s_1 = \mathbf{1}(y') \cap \mathbf{0}(x^*).$$

This implies that

$$l_\tau(\mathbf{1}(y), s_0) < b_{s_0} \quad \text{and} \quad l_\tau(\mathbf{1}(y'), s_1) < b_{s_1}.$$

Hence, according to the coincidence-detection evolving algorithm, we have

$$l_{\tau+1}(\mathbf{1}(y), s_0) = l_\tau(\mathbf{1}(y), s_0) + \sum_{j \in \mathbf{1}(y)} \mathcal{D}_{x(\tau) \to x(\tau+1)} a_{s_0 j} < b_{s_0}$$

and

$$l_{\tau+1}(\mathbf{1}(y'), s_1) = l_\tau(\mathbf{1}(y'), s_1) + \sum_{j \in \mathbf{1}(y')} \mathcal{D}_{x(\tau) \to x(\tau+1)} a_{s_1 j} < b_{s_1},$$

proving (99).

Case 3. $s_0 \in \mathbf{1}(x^*)$ *and* $s_1 \in \mathbf{0}(x^*)$. Then, by (98), we have

$$\mathbf{1}(y) \subsetneq \mathbf{1}(y') \quad \text{and} \quad \mathbf{1}(y') \supsetneq \mathbf{1}(x^*),$$

and

$$s_0 = \mathbf{0}(y) \cap \mathbf{1}(y') \quad \text{and} \quad s_1 = \mathbf{1}(y') \cap \mathbf{0}(x^*).$$

This implies that

$$l_\tau(\mathbf{1}(y), s_0) \geq b_{s_0} \quad \text{and} \quad l_\tau(\mathbf{1}(y'), s_1) < b_{s_1}.$$

Since

$$\begin{aligned}\mathbf{1}(y) \cap \mathbf{1}(x^*) &= (\mathbf{1}(y) \cup \mathbf{0}(y')) \cap \mathbf{1}(x^*)\\ &= \mathbf{1}(x^*) \setminus (\mathbf{0}(y) \cap \mathbf{1}(y'))\\ &= \mathbf{1}(x^*) \setminus s_0\end{aligned}$$

and

$$\begin{aligned}\mathbf{1}(y) \cap \mathbf{0}(x^*) &= (\mathbf{1}(y') \cap \mathbf{0}(x^*)) \setminus (\mathbf{0}(y) \cap \mathbf{1}(y') \cap \mathbf{0}(x^*))\\ &= (\mathbf{1}(y') \cap \mathbf{0}(x^*)) \setminus (s_0 \cap s_1)\\ &= s_1,\end{aligned}$$

it follows from (41) and the coincidence-detection evolving algorithm that

$$\begin{aligned}\sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j} &= \sum_{j\in\mathbf{1}(y)\cap\mathbf{1}(x^*)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j} + \sum_{j\in\mathbf{1}(y)\cap\mathbf{0}(x^*)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j}\\ &= \sum_{j\in\mathbf{1}(x^*), j\neq s_0} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j} + \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 s_1} \geq 0.\end{aligned}$$

Thus

$$l_{\tau+1}(\mathbf{1}(y), s_0) = l_\tau(\mathbf{1}(y), s_0) + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j} \geq b_{s_0}.$$

On the other hand, since $s_1 \in \mathbf{0}(x^*)$, it follows from the coincidence-detection evolving algorithm that

$$l_{\tau+1}(\mathbf{1}(y'), s_1) = l_\tau(\mathbf{1}(y'), s_1) + \sum_{j\in\mathbf{1}(y')} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_1 j} < b_{s_1},$$

proving (99).

Case 4. $s_0 \in \mathbf{0}(x^*)$ *and* $s_1 \in \mathbf{1}(x^*)$. Then, by (98), we have

$$\mathbf{1}(y) \supsetneq \mathbf{1}(y') \quad \text{and} \quad \mathbf{1}(y') \subsetneq \mathbf{1}(x^*)$$

and

$$s_0 = \mathbf{1}(y) \cap \mathbf{0}(y') \quad \text{and} \quad s_1 = \mathbf{0}(y') \cap \mathbf{1}(x^*).$$

This implies that

$$l_\tau(\mathbf{1}(y), s_0) < b_{s_0} \quad \text{and} \quad l_\tau(\mathbf{1}(y'), s_1) \geq b_{s_1}.$$

Hence, according to the coincidence-detection evolving algorithm, we have

$$l_{\tau+1}(\mathbf{1}(y), s_0) = l_\tau(\mathbf{1}(y), s_0) + \sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_0 j} < b_{s_0}$$

and

$$l_{\tau+1}(\mathbf{1}(y'), s_1) = l_\tau(\mathbf{1}(y'), s_1) + \sum_{j\in\mathbf{1}(y')} \mathcal{D}_{x(\tau)\to x(\tau+1)} a_{s_1 j} \geq b_{s_1},$$

proving (99). This implies that

$$Pro_{A(\tau+1)}(y, x^*) \leq 2,$$

completing the proof of (97).

With the notation and the arguments above, we describe the mathematical feature of neural synchrony as follows.

THEOREM 5. *If x^* is a synchronization state, then*

(i) $a_{ij}(t+1) \geq a_{ij}(t)$ *for all* $t \geq T$ *if* $i, j \in \mathbf{1}(x^*)$*; otherwise* $a_{ij}(t+1) \leq a_{ij}(t)$ *for all* $t \geq T$*;*

(ii) $\{y;\ Pro_{A(T)}(y, x^*) \leq 2\} \subset \{y;\ Pro_{A(T+1)}(y, x^*) \leq 2\} \subset \cdots.$

8. Determination of the size of neural synchrony. We have shown in Theorem 2 that, after a finite number of time steps, synchronously firing neural groups emerge. But it is possibly the case that the dynamics can evolve to zero activity or tend to synchronize the entire network. In that case it seems to be irrelevant to the brain function. This raises a question: *How is the size of the synchronous group determined?*

To solve this, we need a criterion for predicting neuronal activty states posterior to each time t. We show that for any given $t = 0, 1, \ldots$ and $x(t), A(t), s(t)$ such that

$$x(t+1) = H_{A(t),s(t)}(x(t))$$

by (3) and (4), the regions of all possible neuronal activity states generated at times $t+2$ and $t+3$ can be dominated by the regions of states generated according to the former $A(t)$ and $x(t+1)$.

THEOREM 6. *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Let the discrete flow $x(t)$ iterate asynchronously and let $A(t)$ satisfy the condition of assembling coordination described in Theorem* 2*. Then for each $t = 0, 1, \ldots,$ the inclusion holds:*

$$\{y;\ Pro_{A(t+1)}(x(t+1), y) \leq 1\} \subset \{y;\ Pro_{A(t)}(x(t+1), y) \leq 1\}. \tag{100}$$

If, in addition, the plasticity parameters satisfy (41)*, then*

$$\{y;\ Pro_{A(t+2)}(x(t+2), y) \leq 1\} \subset \{y;\ Pro_{A(t)}(x(t+2), y) \leq 1\} \tag{101}$$

and

$$\{y;\ Pro_{A(t+2)}(x(t+2), y) \leq 1\} \subset \{y;\ Pro_{A(t)}(x(t+1), y) \leq 2\}. \tag{102}$$

Proof. Let $x(0)$ be any initial neuronal active state in $\{0,1\}^n$, and let $x(t)$ iterate asynchronously, guided by the dynamics (3), (4) and the coincidence-detection evolving algorithm. To prove (100), it suffices to show that for each $t = 0, 1, \ldots,$ if there exists y with $Pro_{A(t+1)}(x(t+1), y) = 1$, then $Pro_{A(t)}(x(t+1), y) = 1$. Fixing t and y, we split the arguments into two cases.

Case 1. $\mathbf{1}(x(t+1)) \subsetneq \mathbf{1}(y)$. Let $s = \mathbf{0}(x(t+1)) \cap \mathbf{1}(y)$. Since

$$Pro_{A(t+1)}(x(t+1), y) = 1,$$

we have

$$H_{A(t+1),s}(x(t+1)) = y.$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+1)), s) \geq b_s. \tag{103}$$

Since

$$l_{t+1}(\mathbf{1}(x(t+1)), s) = l_t(\mathbf{1}(x(t+1)), s) + \sum_{j \in \mathbf{1}(x(t+1))} \mathcal{D}_{x(t) \to x(t+1)} a_{sj}$$

and $s \in \mathbf{0}(x(t+1))$, it follows from (103) and the coincidence-detection evolving algorithm that

$$l_t(\mathbf{1}(x(t+1)), s) \geq l_{t+1}(\mathbf{1}(x(t+1)), s) \geq b_s.$$

This implies that

$$H_{A(t),s}(x(t+1)) = y,$$

and hence

$$Pro_{A(t)}(x(t+1), y) = 1.$$

Case 2. $\mathbf{1}(x(t+1)) \supsetneq \mathbf{1}(y)$. Let $s = \mathbf{1}(x(t+1)) \cap \mathbf{0}(y)$. Since

$$Pro_{A(t+1)}(x(t+1), y) = 1,$$

we have

$$H_{A(t+1),s}(x(t+1)) = y.$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+1)), s) < b_s. \tag{104}$$

Since

$$l_{t+1}(\mathbf{1}(x(t+1)), s) = l_t(\mathbf{1}(x(t+1)), s) + \sum_{j \in \mathbf{1}(x(t+1))} \mathcal{D}_{x(t) \to x(t+1)} a_{sj}$$

and $s \in \mathbf{1}(x(t+1))$, it follows from (104) and the coincidence-detection evolving algorithm that

$$l_t(\mathbf{1}(x(t+1)), s) \leq l_{t+1}(\mathbf{1}(x(t+1)), s) < b_s.$$

This implies that

$$H_{A(t),s}(x(t+1)) = y,$$

and hence

$$Pro_{A(t)}(x(t+1),y)=1,$$

proving (100).

To prove (101) and (102), we have to claim that

$$\{y;\ Pro_{A(t+1)}(x(t+2),y)\le 1\}\subset\{y;\ Pro_{A(t)}(x(t+2),y)\le 1\}. \tag{105}$$

So, together with (100), we have

$$\begin{aligned}\{y;\ Pro_{A(t+2)}(x(t+2),y)\le 1\}&\subset\{y;\ Pro_{A(t+1)}(x(t+2),y)\le 1\}\\ &\subset\{y;\ Pro_{A(t)}(x(t+2),y)\le 1\}.\end{aligned}$$

Also, by (100), we get

$$Pro_{A(t)}(x(t+1),x(t+2))\le Pro_{A(t+1)}(x(t+1),x(t+2))\le 1.$$

Thus for each y with $Pro_{A(t+2)}(x(t+2),y)\le 1$, we have

$$Pro_{A(t)}(x(t+1),y)\le Pro_{A(t)}(x(t+1),x(t+2))+Pro_{A(t)}(x(t+2),y)\le 2.$$

To prove (105), it suffices to show that if there exists y with

$$Pro_{A(t+1)}(x(t+2),y)=1,$$

then

$$Pro_{A(t)}(x(t+2),y)=1.$$

Case 1. $\mathbf{1}(x(t+2))\subsetneq\mathbf{1}(y)$. Let $s=\mathbf{0}(x(t+2))\cap\mathbf{1}(y)$. Since

$$Pro_{A(t+1)}(x(t+2),y)=1,$$

we have

$$H_{A(t+1),s}(x(t+2))=y.$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+2)),s)\ge b_s. \tag{106}$$

Since

$$l_{t+1}(\mathbf{1}(x(t+2)),s)=l_t(\mathbf{1}(x(t+2)),s)+\sum_{j\in\mathbf{1}(x(t+2))}\mathcal{D}_{x(t)\to x(t+1)}a_{sj},$$

it follows from (106) and the coincidence-detection evolving algorithm that

$$l_t(\mathbf{1}(x(t+2)),s)\ge l_{t+1}(\mathbf{1}(x(t+2)),s)\ge b_s \tag{107}$$

if $s\in\mathbf{0}(x(t+1))$. Indeed, if $s\notin\mathbf{0}(x(t+1))$, then, according to the asynchronous iteration of $x(t)$, we have

$$s=\mathbf{1}(x(t+1))\cap\mathbf{0}(x(t+2))$$

and

$$H_{A(t+1),s}(x(t+1)) = x(t+2).$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+1)), s) < b_s,$$

and hence, applying the assembling coordination (10) to $U = s$, $t_* = 0$ and $t^* = t+1$, we get

$$\begin{aligned} l_{t+1}(\mathbf{1}(x(t+2)), s) &= l_{t+1}(\mathbf{1}(x(t+1)), s) - a_{ss}(t+1) \\ &< b_s - a_{ss}(t+1) \leq b_s, \end{aligned}$$

contradicting (106). Thus, by (107), we conclude that

$$H_{A(t),s}(x(t+2)) = y,$$

and hence

$$Pro_{A(t)}(x(t+2), y) = 1.$$

Case 2. $\mathbf{1}(x(t+2)) \supsetneqq \mathbf{1}(y)$. Let $s = \mathbf{1}(x(t+2)) \cap \mathbf{0}(y)$. Since

$$Pro_{A(t+1)}(x(t+2), y) = 1,$$

we have

$$H_{A(t+1),s}(x(t+2)) = y.$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+2)), s) < b_s. \tag{108}$$

Further we have $s \in \mathbf{1}(x(t+1))$. Indeed, if $s \notin \mathbf{1}(x(t+1))$, then, according to the asynchronous iteration of $x(t)$, we have

$$s = \mathbf{0}(x(t+1)) \cap \mathbf{1}(x(t+2))$$

and

$$H_{A(t+1),s}(x(t+1)) = x(t+2).$$

This implies that

$$l_{t+1}(\mathbf{1}(x(t+1)), s) \geq b_s.$$

Applying the assembling coordination (10) to $U = s$, $t_* = 0$, and $t^* = t+1$, we get

$$\begin{aligned} l_{t+1}(\mathbf{1}(x(t+2)), s) &= l_{t+1}(\mathbf{1}(x(t+1)), s) + a_{ss}(t+1) \\ &\geq b_s + a_{ss}(t+1) \geq b_s, \end{aligned}$$

contradicting (108). Thus, by the coincidence-detection evolving algorithm, we have

$$\sum_{j \in \mathbf{1}(x(t+2))} \mathcal{D}_{x(t) \to x(t+1)} a_{sj} \geq 0 \tag{109}$$

if $\mathbf{1}(x(t+2)) \subset \mathbf{1}(x(t+1))$, and together with (41) and applying the assembling coordination (10) to $U = s$, $t_* = 0$, and $t^* = t+1$, we see that

$$\begin{aligned}\sum_{j\in\mathbf{1}(x(t+2))} \mathcal{D}_{x(t)\to x(t+1)} a_{sj} = & \sum_{j\in\mathbf{1}(x(t+1)),j\neq s} \mathcal{D}_{x(t)\to x(t+1)} a_{sj} + \mathcal{D}_{x(t)\to x(t+1)} a_{ss} \\ & + \mathcal{D}_{x(t)\to x(t+1)} a_{s\mathbf{0}(x(t+1))\cap\mathbf{1}(x(t+2))} \geq 0\end{aligned} \tag{110}$$

if $\mathbf{1}(x(t+2)) \not\subset \mathbf{1}(x(t+1))$. Since

$$l_{t+1}(\mathbf{1}(x(t+2)), s) = l_t(\mathbf{1}(x(t+2)), s) + \sum_{j\in\mathbf{1}(x(t+2))} \mathcal{D}_{x(t)\to x(t+1)} a_{sj},$$

we conclude from (108), (109), and (110) that

$$l_t(\mathbf{1}(x(t+2)), s) \leq l_{t+1}(\mathbf{1}(x(t+2)), s) < b_s.$$

This implies

$$H_{A(t),s}(x(t+2)) = y,$$

and hence

$$Pro_{A(t)}(x(t+2), y) = 1,$$

proving (105), and the theorem follows. □

Theorem 6 yields the determination of the size of neural synchrony.

THEOREM 7. *Consider the evolutionary network of n coupled neurons subject to the dynamics* (3) *and* (4) *and obeying the coincidence-detection evolving algorithm. Let the discrete flow $x(t)$ iterate asynchronously, the evolutionary coupling state $A(t)$ satisfy the condition of assembling coordination described in Theorem* 2, *and the plasticity parameters satisfy* (41). *Let Ω be a nonempty subset of $\{0,1\}^n$ with $0 \in \Omega$ and $k = 1, 2$. Suppose for every $y \in \{0,1\}^n$, $z \in \Omega$ with*

$$Pro_{A(t)}(x(t+1), y) = k-1 \quad \textit{and} \quad Pro_{A(t)}(y, z) = 1,$$

the plasticity parameters $\mathcal{D}_{x(t)\to x(t+1)} a_{ij}$ satisfy the conditions

$$\sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(t)\to x(t+1)} a_{sj} \geq b_s - \sum_{j\in\mathbf{1}(y)} a_{sj}(t) \ \textit{ whenever } s = \mathbf{1}(x(t+1)) \cap \mathbf{1}(y) \cap \mathbf{0}(z), \tag{111}$$

$$\sum_{j\in\mathbf{1}(y)} \mathcal{D}_{x(t)\to x(t+1)} a_{sj} < b_s - \sum_{j\in\mathbf{1}(y)} a_{sj}(t) \ \textit{ whenever } s = \mathbf{0}(x(t+1)) \cap \mathbf{0}(y) \cap \mathbf{1}(z). \tag{112}$$

If $Pro_{A(0)}(x(0), \Omega) > k$, then

$$Pro_{A(t)}(x(t), \Omega) > k \ \textit{ for } t = 1, 2, \ldots, \tag{113}$$

and

$$\lim_{t\to\infty} x(t) \notin \Omega. \tag{114}$$

Proof. It suffices to prove (113), because (114) follows immediately from (113) and Theorem 2. Fix $k = 1, 2$, and let $x(0)$ and $A(0)$ be the initial states satisfying

$$Pro_{A(0)}(x(0), \Omega) > k.$$

Having chosen $x(0), x(1), \ldots, x(\tau)$ and $A(0), A(1), \ldots, A(\tau)$ with

$$Pro_{A(t)}(x(t), \Omega) > k \quad \text{for } t = 0, 1, \ldots, \tau,$$

we see from (3) and (4) that $x(\tau + 1)$ satisfies

$$Pro_{A(\tau)}(x(\tau + 1), \Omega) \geq k.$$

Then, by (111) and (112), we see that the construction of $A(\tau + 1)$ satisfies

$$Pro_{A(\tau+1)}(x(\tau + 1), \Omega) \neq k. \tag{115}$$

Indeed, if there exists $z \in \Omega$ such that $Pro_{A(\tau+1)}(x(\tau + 1), z) = k$, then we can find $s_0, s_1, \ldots, s_{k-1} \in \{1, 2, \ldots, n\}$ such that

$$H_{A(\tau+1),s_{k-1}} \circ H_{A(\tau+1),s_{k-2}} \circ \cdots \circ H_{A(\tau+1),s_0}(x(\tau + 1)) = z.$$

Put

$$\begin{aligned} z(0) &= x(\tau), \\ z(1) &= x(\tau + 1), \\ z(2) &= H_{A(\tau+1),s_0}(z(1)), \\ &\vdots \\ z(k + 1) &= H_{A(\tau+1),s_{k-1}}(z(k)) \end{aligned}$$

and

$$W(0) = A(\tau), W(1) = W(2) = \cdots = W(k) = A(\tau + 1).$$

Then $z(0), z(1), \ldots, z(k+1)$ are mutually distinct. Since $Pro_{W(1)}(z(1), z(k)) = k-1$, it follows from Theorem 6 that

$$Pro_{W(0)}(z(1), z(k)) = k - 1.$$

Since $Pro_{W(k)}(z(k), z(k + 1)) = 1$, it follows from Theorem 6 that

$$Pro_{W(0)}(z(k), z(k + 1)) = 1.$$

Note that $z(k + 1) = z \in \Omega$, and

$$\begin{aligned} \mathbf{1}(z(k)) \cap \mathbf{0}(z(k + 1)) &\subset \mathbf{1}(z(k - 1)) \cap \mathbf{1}(z(k)), \\ \mathbf{0}(z(k)) \cap \mathbf{1}(z(k + 1)) &\subset \mathbf{0}(z(k - 1)) \cap \mathbf{0}(z(k)), \end{aligned}$$

so by (111) and (112), we have

$$\sum_{j \in \mathbf{1}(z(k))} a_{sj}(\tau + 1) \geq b_s \quad \text{if } s = \mathbf{1}(z(k)) \cap \mathbf{0}(z(k + 1))$$

and

$$\sum_{j\in\mathbf{1}(z(k))} a_{sj}(\tau+1) < b_s \quad \text{if } s = \mathbf{0}(z(k)) \cap \mathbf{1}(z(k+1)),$$

in contradiction to

$$z(k+1) = H_{A(\tau+1),s_{k-1}}(z(k)).$$

Now, according to the inductive assumption

$$Pro_{A(\tau)}(x(\tau+1),\Omega) \geq k$$

and the inclusion derived from Theorem 6,

$$\{y;\ Pro_{A(\tau+1)}(x(\tau+1),y) \leq k-1\} \subset \{y;\ Pro_{A(\tau)}(x(\tau+1),y) \leq k-1\},$$

we have

$$Pro_{A(\tau+1)}(x(\tau+1),\Omega) \geq k.$$

Combining this with (115) implies that

$$Pro_{A(\tau+1)}(x(\tau+1),\Omega) > k,$$

completing the inductive proof of (113) and the proof of Theorem 7 (see Figure 6 for an illustration). ☐

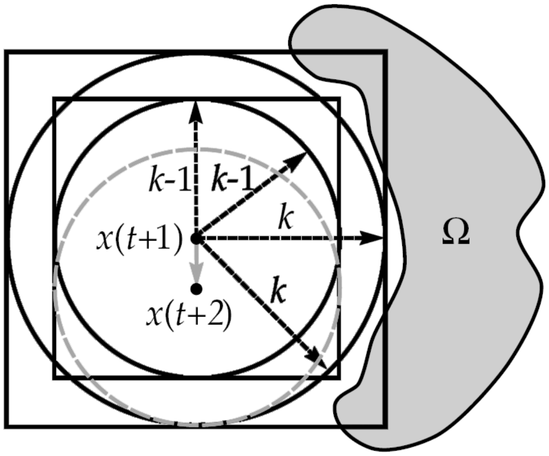


FIG. 6. *The regime of plasticity parameters described in Theorem* 7 *reveals an inductive scheme for the construction of the evolutionary coupling strengths so that the synchronously firing groups of neurons can be determined by a nonempty subset* Ω *of* $\{0,1\}^n$*. For each* $t = 0,1,\ldots$ *and each choice of* $A(t)$*, if* $Pro_{A(t)}(x(t+1),\Omega) \geq k$ *(illustrated by the bigger square that meets* Ω *and the smaller square that doesn't meet* Ω*), then we need only focus on the choice of* $A(t+1)$ *such that* $Pro_{A(t+1)}(x(t+1),\Omega) \neq k$*, and according to Theorem* 6*, it is guaranteed that* $Pro_{A(t+1)}(x(t+1),\Omega) > k$ *(illustrated by both the bigger and the smaller black circles that don't meet* Ω*). Therefore, when* $x(t+1)$ *iterates (the gray arrow), we have* $Pro_{A(t+1)}(x(t+2),\Omega) \geq k$ *(the gray circle centered at* $x(t+2)$ *is a translation of the smaller black circle centered at* $x(t+1)$ *and it doesn't meet* Ω*), yielding the inductive scheme.*

For a specialized region

$$\Omega = \{x;\ x_1 + \cdots + x_n = l_1\} \cup \{x;\ x_1 + \cdots + x_n = l_2\},$$

where $0 < l_1 \ll l_2 < n$, if the initial state $x(0)$ satisfies

$$l_1 + k < x_1(0) + \cdots + x_n(0) < l_2 - k$$

for $k = 1, 2$ and the discrete flow $x(t)$ fulfills the parameter regime of Theorem 7, then Theorem 7 implies that the evolutionary network will eventually evolve to a state of a synchronously firing group of neurons whose size is greater than l_1 and less than l_2. This implies that, with the Hebbian synaptic plasticity, there are regimes of plasticity parameters such that global synchrony or zero activity cannot occur in the dynamics of the evolutionary network.

9. Conclusions.

We describe a model of evolutionary neural networks and unravel its meta-complication of nodal and coupling dynamics. We show that there are time- and activity-dependent nodal-and-coupling changes in the dynamics of the evolutionary network and prove that the dynamics will eventually result in a state of synchronization if those changes are based on Hebbian synaptic plasticity.

Furthermore, we study the stability problem of neural synchrony. We introduce the proximal number to quantify the disturbance of neuronal activity states and reveal the phenomenon of local absorption in a network's dynamics. We show that, underlying Hebbian synaptic plasticity, small perturbations of the initial neuronal active state can cause only small variations of the discrete flow which iterates to a state of neural synchrony.

The combined effect of neural synchrony and its stability shows that neural synchrony is not only to spark tighter connections between neurons but also to generate a monotonically increasing region of states for initializing the stability process. This implies the robust stability of neural synchrony.

We also show that, with Hebbian synaptic plasticity, there is a criterion for predicting neuronal activity states while the evolutionary network evolves. It represents an inductive scheme for the determination of the size of synchronous groups.

Our model of neural processing reflects, in a deep mathematical sense, that plasticity emerges not only as a source of computational power but a source of spontaneous order as well. The extent of plasticity, synchrony, and stability, and the laws it depicts, may find use across many levels in nature and society, appealing to the understanding of the growth dynamics of assemblies, clusters, or communities to interpret the organization of hierarchical architecture in complex systems.

REFERENCES

[1] P. Adams, *Hebb and Darwin*, J. Theoret. Biol., 195 (1998), pp. 419–438.

[2] J. A. Anderson, *A simple neural network generating an interactive memory*, Math. Biosci., 14 (1972), pp. 197–220.

[3] J. A. Anderson and E. Rosenfeld, *Neurocomputing*, The MIT Press, Cambridge, MA, 1988.

[4] T. Ando and M.-H. Shih, *Simultaneous contractibility*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 487–498.

[5] W. Aspray and A. Burks, *Papers of John von Neumann on Computing and Computer Theory*, The MIT Press, Cambridge, MA, 1986.

[6] J. F. Brons and C. D. Woody, *Long-term changes in excitability of cortical neurons after Pavlovian conditioning and extinction*, J. Neurophysiol., 44 (1980), pp. 605–615.

[7] S. R. Cajal, *La fine structure des centres nerveux*, Proc. Roy. Soc. London, 55 (1894), pp. 444–468.

[8] G. Chappell and J. Taylor, *The temporal Kohonen map*, Neural Netw., 6 (1993), pp. 441–445.

[9] M. A. Cohen and S. Grossberg, *Absolute stability of global pattern formation and parallel memory storage by competitive neural networks*, IEEE Trans. Systems Man Cybernet., SMC-13 (1983), pp. 815–826.

[10] G. Daoudal and D. Debanne, *Long-term plasticity of intrinsic excitability: Learning rules and mechanisms*, Learn. Mem., 10 (2003), pp. 456–465.

[11] S. Dehaene, M. Kerszberg, and J.-P. Changeux, *A neuronal model of a global workspace in effortful cognitive tasks*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 14529–14534.

[12] N. S. Desai, L. C. Rutherford, and G. G. Turrigiano, *Plasticity in the intrinsic excitability of cortical pyramidal neurons*, Nature Neurosci., 2 (1999), pp. 515–520.

[13] A. K. Engel, P. Fries, and W. Singer, *Dynamic predictions: Oscillations and synchrony in top-down processing*, Nat. Rev. Neurosci., 2 (2001), pp. 704–716.

[14] J. C. Forte and G. Pagés, *On the a.s. convergence of the Kohonen algorithm with a general neighborhood function*, Ann. Appl. Probab., 5 (1995), pp. 1177–1216.

[15] S. Grossberg, *Adaptive pattern classification and universal recoding,* I: *Parallel development and coding of neural feature detectors*, Biol. Cybernet., 23 (1976), pp. 121–134.

[16] S. Grossberg, *Adaptive pattern classification and universal recoding,* II: *Feedback, expectation, olfaction, and illusions*, Biol. Cybernet., 23 (1976), pp. 187–202.

[17] S. Grossberg, *Nonlinear neural networks: Principles, mechanisms, and architectures*, Neural Netw., 1 (1988), pp. 17–61.

[18] S. Grossberg, *How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex*, Spat. Vis., 12 (1999), pp. 163–185.

[19] B. Hammer, A. Micheli, A. Sperduti, and M. Strickert, *Recursive self-organizing network models*, Neural Netw., 17 (2004), pp. 1061–1085.

[20] K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsáki, *Organization of cell assemblies in the hippocampus*, Nature, 424 (2003), pp. 552–556.

[21] D. O. Hebb, *The Organization of Behavior*, Wiley, New York, 1949.

[22] J. A. Hertz, A. S. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, New York, 1991.

[23] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci. USA, 79 (1982), pp. 2554–2558.

[24] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using neatest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[25] Y. Kamp and M. Hasler, *Recursive Neural Networks for Associative Memory*, John Wiley & Sons, New York, 1990.

[26] M. Klein, B. Hochner, and E. R. Kandel, *Facilitatory transmitters and cAMP can modulate accommodation as well as transmitter release in Aplysia sensory neurons. Evidence for parallel processing in a single cell*, Proc. Natl. Acad. Sci. USA, 83 (1986), pp. 7994–7998.

[27] T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biol. Cybernet., 43 (1982), pp. 59–69.

[28] T. Kohonen, *Self-organization and Associative Memory*, Springer, Berlin, 1989.

[29] T. Kohonen, *Self-organizing Maps*, Springer, Berlin, 1995.

[30] J. Lücke, *Hierarchical self-organization of minocolumnar receptive fields*, Neural Netw., 17 (2004), pp. 1377–1389.

[31] J. Lücke and C. V. D. Malsburg, *Rapid processing and unsupervised learning in a model of the cortical macrocolumn*, Neural Comput., 16 (2004), pp. 501–533.

[32] E. Marder, L. F. Abbott, G. G. Turrigiano, Z. Liu, and J. Golowasch, *Memory from the dynamics of intrinsic membrane currents*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 13481–13486.

[33] W. S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bull. Math. Biophys., 5 (1943), pp. 115–133.

[34] B. Milner, L. R. Squire, and E. R. Kandel, *Cognitive neuroscience and the study of memory*, Neuron, 20 (1998), pp. 445–468.

[35] M. Minsky, *Computation: Finite and Infinite Machines*, Prentice-Hall, New York, 1967.

[36] M. A. L. Nicolelis, E. E. Fanselow, and A. A. Ghazanfar, *Hebb's dream: The resurgence of cell assemblies*, Neuron, 19 (1997), pp. 219–221.

[37] J. Principe, N. Euliano, and S. Garani, *Principles and networks for self-organization in space–time*, Neural Netw., 15 (2002), pp. 1069–1083.

[38] N. Rochester, J. H. Holland, L. H. Haibt, and W. L. Duda, *Tests on a cell assembly theory of the action of the brain, using a large digital computer*, IRE Trans. Inf. Theory, 2 (1956), pp. 80–93.

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, Nature, 323 (1986), pp. 533–536.

[40] B. Schechter, *How the brain gets rhythm*, Science, 274 (1996), pp. 339–340.

[41] T. J. Sejnowski, *The book of Hebb*, Neuron, 24 (1999), pp. 773–776.

[42] C. S. Sherrington, *The Integrative Action of the Nervous System*, Yale University Press, New Haven, CT, 1906.

[43] M.-H. Shih, *Simultaneous Schur stability*, Linear Algebra Appl., 287 (1999), pp. 323–336.

[44] M.-H. Shih and J.-L. Dong, *A combinatorial analogue of the Jacobian problem in automata networks*, Adv. in Appl. Math., 34 (2005), pp. 30–46.

[45] M.-H. Shih and C.-T. Pang, *Simultaneous Schur stability of interval matrices*, Automatica, in press (2008).

[46] A. M. Sillito, H. E. Jones, G. L. Gerstein, and D. C. West, *Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex*, Nature, 369 (1994), pp. 479–482.

[47] A. P. Sripati and K. O. Johnson, *Dynamic gain changes during attentional modulation*, Neural Comput., 18 (2006), pp. 1847–1867.

[48] A. V. Stein, C. Chiang, and P. König, *Top-down processing mediated by interareal synchronization*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 14748–14753.

[49] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[50] G. Tononi and G. M. Edelman, *Consciousness and complexity*, Science, 282 (1998), pp. 1846–1851.

[51] G. G. Turrigiano, L. F. Abbott, and E. Marder, *Activity-dependent changes in the intrinsic properties of cultured neurons*, Science, 264 (1994), pp. 974–977.

[52] F. Varela, J.-P. Lachaux, E. Rodriguez, and J. Martinerie, *The brainweb: Phase synchronization and large-scale integration*, Nat. Rev. Neurosci., 2 (2001), pp. 229–239.

[53] T. Voegtlin, *Recursive self-organizing maps*, Neural Netw., 15 (2002), pp. 979–991.

[54] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

[55] B. Widrow and M. E. Hoff, *Adaptive switching circuits*, IRE WESCON Convention Record, New York, 1960, pp. 96–104.

[56] D. J. Willshaw and C. V. D. Malsburg, *How patterned neural connections can be set up by self-organizations*, Proc. Roy. Soc. London Ser. B, 194 (1976), pp. 431–445.

[57] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, *Non-holographic associative memory*, Nature, 222 (1969), pp. 960–962.

[58] A. Yazdanbakhsh and S. Grossberg, *Fast synchronization of perceptual grouping in laminar visual cortical circuits*, Neural Netw., 17 (2004), pp. 707–718.

[59] W. Zhang and D. J. Linden, *The other side of the engram: Experience-driven changes in neuronal intrinsic excitability*, Nature Rev. Neurosci., 4 (2003), pp. 885–900.

SIAM J. APPL. MATH.
Vol. 69, No. 4, pp. 1162–1173

THE STRONGLY CONFINED SCHRÖDINGER–POISSON SYSTEM FOR THE TRANSPORT OF ELECTRONS IN A NANOWIRE*

NAOUFEL BEN ABDALLAH†, FRANCOIS CASTELLA‡, FANNY DELEBECQUE-FENDT‡, AND FLORIAN MÉHATS‡

Abstract. We study the limit of the three-dimensional Schrödinger–Poisson system with a singular perturbation, to model a quantum electron gas that is strongly confined near an axis. For well-prepared data, which are polarized on the ground space of the transversal Hamiltonian, the resulting model is the cubic defocusing nonlinear Schrödinger equation. Our main tool is a refined analysis of the Poisson kernel when acting on strongly confined densities. In that direction, an appropriate scaling of the initial data is required, to avoid divergent integrals when the gas concentrates on the axis.

Key words. Schrödinger–Poisson system, asymptotic analysis, singular perturbation, quantum transport, nanowire, nanoelectronics

AMS subject classifications. 35Q40, 35Q55, 35B40, 35B25, 82D37

DOI. 10.1137/080715950

1. Introduction.

1.1. The physical problem and the singularly perturbed system.

Artificially confined structures are now routinely realized in the nanoelectronic industry, and the functioning of many electronic devices is based on the transport of charged particles which are bounded in transversal directions; see, e.g., [AFS, Bas, FG, VW]. The confinement can be typically monodimensional, like in quantum wells where two directions remain for the transport, or bidimensional, like in quantum wires where the transport is in dimension one. In this work we are interested in the second case, and this paper is devoted to the rigorous derivation of a dynamic one-dimensional quantum model with space-charge effects describing the transport of electrons confined in a nanowire. Compared with some previous works that treat problems of quantum confinement, as, for instance, [CDR] about an asymptotic model for two particles on the surface of a nanotube (see also references therein), our concern here is to deal more specifically with the nonlinear Poisson term.

Our strategy is inspired by that in [BAMP, BMSW, BCM] and consists of an asymptotic analysis of the three-dimensional Schrödinger–Poisson system (or Hartree system), that will be referred to as our "starting model," with a singular perturbation modeling a strong potential confining the electron gas in a wire. The interesting point concerning the reduced model obtained in the limit is that the nonlinearity describing space-charge effects is now *localized*, this reduced model taking the form of a cubic defocusing nonlinear Schrödinger (NLS) equation.

Let us describe the starting model. The space variable is written (x, z_1, z_2), where $x \in \mathbb{R}$ is the direction in which the electron gas is transported free from any external

*Received by the editors February 18, 2008; accepted for publication (in revised form) September 22, 2008; published electronically January 30, 2009. This work partially supported by the ANR project ANR-07-BLAN-0236 QUATRAIN "Quantum transport in Nanostructures."

http://www.siam.org/journals/siap/69-4/71595.html

†Institut de Mathématiques, Université Paul Sabatier, Toulouse 3, France (naoufel@math.univ-toulouse.fr).

‡IRMAR, Université Rennes 1, France (francois.castella@univ-rennes1.fr, fanny.fendt@univ-rennes1.fr, florian.mehats@univ-rennes1.fr).

force and $z = (z_1, z_2) \in \mathbb{R}^2$ are the confined directions. We consider the following singularly perturbed Schrödinger–Poisson system:

$$i\partial_t \Psi^\varepsilon = -\Delta\Psi^\varepsilon + \frac{1}{\varepsilon^2} V_c\left(\frac{z}{\varepsilon}\right)\Psi^\varepsilon + \mathbb{V}^\varepsilon \Psi^\varepsilon, \tag{1.1}$$

$$\mathbb{V}^\varepsilon = \frac{1}{4\pi r} * \left(|\Psi^\varepsilon|^2\right). \tag{1.2}$$

The unknown in this system is the pair $(\Psi^\varepsilon, \mathbb{V}^\varepsilon)$ made of the electronic wave function Ψ^ε and the self-consistent potential $\mathbb{V}^\varepsilon$ due to space charge effects, written here as a convolution with the Poisson kernel. We use the notation $r(x,z) = \sqrt{x^2 + |z|^2}$. The main modeling assumption is that a strong external potential is applied to the gas, written here $\frac{1}{\varepsilon^2} V_c(\frac{z}{\varepsilon})$, where $V_c(z)$ is a prescribed function satisfying the following assumption.

Assumption 1.1. *The function $V_c : \mathbb{R}^2 \mapsto \mathbb{R}$ belongs to $L^2_{loc}(\mathbb{R}^2)$, and there exists $\alpha > 0$ and $C > 0$ such that*

$$V_c(z) \geq C\,|z|^\alpha.$$

The crucial assumption here is the growth at infinity, which determines the strength of the confinement. The parameter $\varepsilon \in (0,1)$ is the scaled thickness of the electron gas. As we will see after a rescaling in the next section, the normalization term $\frac{1}{\varepsilon^2}$ is natural in order to balance the strong external potential with the Laplace operator in the z variable.

This paper studies the asymptotic behavior of $(\Psi^\varepsilon, \mathbb{V}^\varepsilon)$ as ε goes to zero. Of course, an initial datum $\Psi^\varepsilon(0,x,z)$ needs to be prescribed for (1.1), whose specific form is made precise in the next section.

1.2. Scaling of the initial data and formal limit. In this section, we derive heuristically the asymptotic model satisfied by the solution of (1.1)–(1.2) as ε goes to zero. Precise and rigorous statements will be made in the next section. Let us introduce the following notation for averages upon the transversal variables:

$$\langle f \rangle = \int_{\mathbb{R}^2} f(z)dz.$$

The singular term $\frac{1}{\varepsilon^2} V_c(\frac{z}{\varepsilon})$ in the Schrödinger equation (1.1) induces a concentration of the density on the axis $z = 0$. We expect that, as $\varepsilon \to 0$, the density takes the form of a line density multiplied by a delta function:

$$|\Psi^\varepsilon(t,x,z)|^2 \sim \left\langle |\Psi^\varepsilon(t,x,\cdot)|^2 \right\rangle \delta(z). \tag{1.3}$$

The crucial point is the consequence of (1.3) on the self-consistent potential. Indeed, we can prove (see Proposition 2.1) that, near the axis $z = 0$, the solution of (1.2) looks like

$$\mathbb{V}^\varepsilon(t,x,z) \sim -\frac{1}{2\pi} \log\varepsilon \left\langle |\Psi^\varepsilon(t,x,\cdot)|^2 \right\rangle. \tag{1.4}$$

This estimate suggests the following choice of initial data: we choose $\Psi^\varepsilon(0,x,z)$ to be small, of order $|\log\varepsilon|^{-1/2}$ (e.g., in $L^2(\mathbb{R}^3)$).

In order to observe the system at the scale of the gas, we work with rescaled space variables, setting

$$\Psi^\varepsilon(t,x,z) = \frac{1}{\varepsilon\sqrt{|\log\varepsilon|}}\,\psi^\varepsilon\left(t,x,\frac{z}{\varepsilon}\right), \qquad \mathbb{V}^\varepsilon(t,x,z) = \frac{1}{|\log\varepsilon|}V^\varepsilon\left(t,x,\frac{z}{\varepsilon}\right).$$

The system in the new unknowns ψ^ε, V^ε becomes

$$i\partial_t\psi^\varepsilon = -\partial_x^2\psi^\varepsilon + \frac{1}{\varepsilon^2}H_z\psi^\varepsilon + \frac{1}{|\log\varepsilon|}V^\varepsilon\psi^\varepsilon\,, \tag{1.5}$$

$$V^\varepsilon = \frac{1}{4\pi r^\varepsilon} * |\psi^\varepsilon|^2, \qquad r^\varepsilon(x,z) = \sqrt{x^2+\varepsilon^2|z|^2}, \tag{1.6}$$

$$\psi^\varepsilon(0,x,z) = \psi_0^\varepsilon(x,z), \tag{1.7}$$

where the Hamiltonian in the z direction is

$$H_z = -\Delta_z + V_c(z).$$

Inserting (1.4) into (1.5), we obtain that, asymptotically, ψ^ε satisfies

$$i\partial_t\psi^\varepsilon = -\partial_x^2\psi^\varepsilon + \frac{1}{\varepsilon^2}H_z\psi^\varepsilon + \frac{1}{2\pi}\left\langle|\psi^\varepsilon|^2\right\rangle\psi^\varepsilon\,, \tag{1.8}$$

$$\psi^\varepsilon(0,x,z) = \psi_0^\varepsilon(x,z). \tag{1.9}$$

This is our reduced model. An elementary remark is that the term $\left\langle|\psi^\varepsilon|^2\right\rangle$ in the nonlinearity does not depend on the z variable. It is thus easy to filter out the oscillations in time induced by the $\frac{1}{\varepsilon^2}H_z$ term. Indeed the function

$$\phi^\varepsilon = e^{itH_z/\varepsilon^2}\,\psi^\varepsilon$$

satisfies the following problem, independent of ε:

$$i\partial_t\phi^\varepsilon = -\partial_x^2\phi^\varepsilon + \frac{1}{2\pi}\left\langle|\phi^\varepsilon|^2\right\rangle\phi^\varepsilon\,, \tag{1.10}$$

$$\phi^\varepsilon(0,x,z) = \psi_0^\varepsilon(x,z), \tag{1.11}$$

where we used the fact that e^{itH_z/ε^2} is an isometry on $L_z^2(\mathbb{R}^2)$, i.e.,

$$\left\langle|\phi^\varepsilon|^2\right\rangle = \left\langle|\psi^\varepsilon|^2\right\rangle.$$

The limit model can be seen as a system of NLS equations in dimension one. To see this, let us introduce the eigenfunctions $(\chi_k(z))_{k\geq 1}$ of the operator H_z and the associated eigenvalues $(E_k)_{k\geq 1}$. Note that Assumption 1.1 implies that the operator H_z is self-adjoint (see, e.g., [RS, Vol. 2, Theorem X.28]) and, defined as a sum of quadratic forms, is an operator with compact resolvent (see [RS, Vol. 4, Theorem XIII.67]). It possesses purely discrete spectrum and a complete set of eigenfunctions.

The reduced model (1.10), (1.11) can be projected on the χ_k's and is equivalent to the system

$$i\partial_t \varphi_k = -\partial_x^2 \varphi_k + \frac{1}{2\pi}\left(\sum_{j=1}^{\infty} |\varphi_j|^2\right)\varphi_k\,, \tag{1.12}$$

$$\varphi_k(0,x) = \int_{\mathbb{R}} \psi_0(x,z)\,\chi_k(z)dz, \qquad k \in \mathbb{N}^*. \tag{1.13}$$

The solution of the rescaled initial problem is then—formally—asymptotically close to

$$\psi^\varepsilon(t,x,z) = \sum_{k=1}^{\infty} e^{-itE_k/\varepsilon^2}\,\varphi_k(t,x)\,\chi_k(z).$$

1.3. Statement of the main result. Let us introduce the energy space

$$\mathcal{H} = \{u \in H^1(\mathbb{R}^3),\ \sqrt{V_c}u \in L^2(\mathbb{R}^3)\}, \tag{1.14}$$

endowed with the norm

$$\|u\|^2_{\mathcal{H}} = \|u\|^2_{H^1(\mathbb{R}^3)} + \|\sqrt{V_c}u\|^2_{L^2(\mathbb{R}^3)} = \|\partial_x u\|^2_{L^2(\mathbb{R}^3)} + \|H_z^{1/2}u\|^2_{L^2(\mathbb{R}^3)}.$$

Note that Assumption 1.1 yields the following control for functions in the energy space:

$$\forall u \in \mathcal{H}, \qquad \int_{\mathbb{R}^3} |z|^\alpha |u|^2\,dxdz \leq C\|u\|^2_{\mathcal{H}}\,. \tag{1.15}$$

Consider for the rescaled starting model (1.5)–(1.7) a sequence of initial data $(\psi_0^\varepsilon)_{\varepsilon>0}$ satisfying the following assumption.

ASSUMPTION 1.2. *The sequence $(\psi_0^\varepsilon)_{\varepsilon>0}$ is uniformly bounded in $\mathcal{H}$ and converges in $L^2(\mathbb{R}^3)$ to a function ψ_0.*

Standard techniques [BM, Caz, IZL, Cas] allow us to prove that for any $\varepsilon \in (0,1)$ the three-dimensional Schrödinger–Poisson system (1.5)–(1.7) admits a unique global weak solution $(\psi^\varepsilon, V^\varepsilon)$ for $t \in \mathbb{R}$ in the energy space. In order to analyze its limit as $\varepsilon \to 0$, let us summarize the available estimates on ψ^ε. The first one is the L^2 estimate. For all t we have

$$\|\psi^\varepsilon(t)\|^2_{L^2(\mathbb{R}^3)} = \|\psi_0^\varepsilon\|^2_{L^2(\mathbb{R}^3)} \leq C. \tag{1.16}$$

Unfortunately (see Proposition 2.1), this estimate alone does not enable us to bound the self-consistent potential, and one needs at least an estimate on the derivative of ψ^ε with respect to x. Let us now examine the second natural estimate for the Schrödinger–Poisson system, namely, the energy estimate. It reads, in rescaled variables:

$$\begin{aligned} &\|\partial_x\psi^\varepsilon(t)\|^2_{L^2(\mathbb{R}^3)} + \frac{1}{\varepsilon^2}\|H_z^{1/2}\psi^\varepsilon(t)\|^2_{L^2(\mathbb{R}^3)} + \frac{1}{|\log\varepsilon|}\left\|V^\varepsilon(t)|\psi^\varepsilon(t)|^2\right\|_{L^1(\mathbb{R}^3)} \\ &\quad = \|\partial_x\psi_0^\varepsilon\|^2_{L^2(\mathbb{R}^3)} + \frac{1}{\varepsilon^2}\|H_z^{1/2}\psi_0^\varepsilon\|^2_{L^2(\mathbb{R}^3)} + \frac{1}{|\log\varepsilon|}\left\|V^\varepsilon(0)|\psi_0^\varepsilon|^2\right\|_{L^1(\mathbb{R}^3)}\,. \end{aligned} \tag{1.17}$$

Multiplying this equation by ε^2, one can deduce a bound for $\|H_z^{1/2}\psi^\varepsilon(t)\|_{L^2(\mathbb{R}^3)}$ (see the beginning of section 3), *but not for* $\|\partial_x\psi^\varepsilon(t)\|_{L^2(\mathbb{R}^3)}$. However, for a certain set of well-prepared initial data it can be easily proved that this quantity is bounded. As it was remarked in [BMSW], it suffices to consider initial data which are polarized on the first eigenmode χ_1 of the transverse Hamiltonian H_z. This leads to the following theorem, which is our main result.

THEOREM 1.3. *Under Assumptions* 1.1 *and* 1.2, *assume, moreover, that the initial data is nearly polarized on the first eigenmode* χ_1 *of* H_z, *associated with the eigenvalues* E_1 *in the sense:*

$$\left\|(H_z - E_1)^{1/2}\,\psi_0^\varepsilon\right\|_{L^2(\mathbb{R}^3)} \le C\,\varepsilon. \tag{1.18}$$

Then there exist $C > 0$ *such that the solution* ψ^ε *of* (1.5)–(1.7) *satisfies*

$$\|\partial_x\psi^\varepsilon(t)\|_{L^2} \le C, \quad \textit{independently of } \varepsilon > 0 \textit{ and } t \in \mathbb{R}, \tag{1.19}$$

and the following convergence result holds, for all $T > 0$*:*

$$\left\|\psi^\varepsilon(t,x,z) - e^{-itE_1/\varepsilon^2}\varphi(t,x)\,\chi_1(z)\right\|_{L^2(\mathbb{R}^3)} \underset{\varepsilon\to 0}{\longrightarrow} 0 \qquad \textit{uniformly on } [-T,T],$$

where $\varphi(t,x)$ *solves the cubic defocusing NLS equation*

$$i\partial_t\varphi = -\partial_x^2\varphi + \frac{1}{2\pi}|\varphi|^2\varphi\,, \qquad \varphi(0,x) = \int_{\mathbb{R}} \psi_0(x,z)\chi_1(z)dz. \tag{1.20}$$

Note that (1.20) is a particular case of the limit model (1.12), (1.13) derived formally in the previous subsection. The keystone of the convergence proof is the L^2 estimate (1.19) of $\partial_x\psi^\varepsilon$. In the general case of initial data bounded in $\mathcal{H}$ but *not* polarized on the first eigenmode, the following partial result can be proved as an easy extension of Theorem 1.3. *Under the assumption that* (1.19) *holds*, the function $e^{itH_z/\varepsilon^2}\psi^\varepsilon$ converges locally uniformly in $L^2(\mathbb{R}^3)$ to the solution ϕ of

$$i\partial_t\phi = -\partial_x^2\phi + \frac{1}{2\pi}\left\langle|\phi|^2\right\rangle\phi\,, \qquad \phi(0,x,z) = \psi_0(x,z).$$

The outline of the paper is the following. In section 2, we give an asymptotic expansion as $\varepsilon \to 0$ of the solution of the rescaled Poisson equation (1.6), for wavefunctions ψ^ε in a suitable functional space. Section 3 is devoted to the proof of Theorem 1.3. As a first step we use the energy estimate for well-prepared data in order to get an estimate of $\|\partial_x\psi^\varepsilon\|_{L^2}$ in that case. We conclude the proof using a stability result for the cubic NLS equation.

2. Approximation of the Poisson kernel.

In this section, we study the convolution with the Poisson kernel when ε is close to zero. We consider the Poisson potential V^ε, after the rescaling $z \mapsto \varepsilon z$, $x \mapsto x$, and let $\varepsilon \to 0$ in (1.6). In order to make a precise statement, let us first recall the definition of the finite part of a singular integral. For $u \in C^{0,\eta}(\mathbb{R}) \cap L^1(\mathbb{R})$, with $\eta \in (0,1)$, we have

$$\begin{aligned}\mathrm{FP}\int_{\mathbb{R}} \frac{u(x')}{|x-x'|}\,dx' &= \lim_{\eta\to 0}\left(\int_{|x-x'|>\eta} \frac{u(x')}{|x-x'|}dx' + 2u(x)\log\eta\right)\\ &= \int_{|x-x'|<1} \frac{u(x')-u(x)}{|x-x'|}dx' + \int_{|x-x'|>1} \frac{u(x')}{|x-x'|}dx'\,.\end{aligned} \tag{2.1}$$

Both quantities are well defined whenever $u \in \mathcal{C}^{0,\eta}(\mathbb{R}) \cap L^1(\mathbb{R})$. Our aim here is to prove the following result.

PROPOSITION 2.1. *Consider ψ in the energy space $\mathcal{H}$ defined by (1.14), and let*

$$G^\varepsilon(\psi) = \int_{\mathbb{R}} \int_{\mathbb{R}^2} \frac{|\psi(x',z')|^2}{\sqrt{(x-x')^2 + \varepsilon^2 |z-z'|^2}}\, dx' dz'.$$

Then we have the following asymptotic expansion,

$$G^\varepsilon(\psi) = -2 \log \varepsilon \langle |\psi(x,\cdot)|^2 \rangle + R_1(\psi) + R_2^\varepsilon(\psi), \tag{2.2}$$

where

$$R_1(\psi) = -2 \int_{\mathbb{R}^2} \log|z-z'| |\psi(x,z')|^2 dz' + 2\log 2 \langle |\psi(x,\cdot)|^2 \rangle + \mathrm{FP} \int_{\mathbb{R}} \frac{\langle |\psi(x',\cdot)|^2 \rangle}{|x-x'|} dx',$$

and for all $u \in \mathcal{H}$ we have

$$\|R_1(\psi) u\|_{L^2} \le C \|\psi\|_{\mathcal{H}}^2 \, \|u\|_{\mathcal{H}} , \qquad \|R_2^\varepsilon(\psi) u\|_{L^2} \le C_\beta \, \varepsilon^\beta \|\psi\|_{\mathcal{H}}^2 \, \|u\|_{\mathcal{H}} \tag{2.3}$$

for all $\beta < \min(1/2, \alpha/2)$, α being defined according to Assumption 1.1.

Proof. Let us first list some useful available estimates deduced from Sobolev embeddings and from (1.15): for all $u \in \mathcal{H}$, we have

$$\|\partial_x u\|_{L^2(\mathbb{R}^3)} + \|u\|_{L^\infty_x L^2_z} + \|\partial_z u\|_{L^2(\mathbb{R}^3)} + \|(1+|z|^{\alpha/2}) u\|_{L^2(\mathbb{R}^3)} \le C \, \|u\|_{\mathcal{H}}. \tag{2.4}$$

Let us now decompose

$$\begin{aligned} G^\varepsilon(\psi) &= \int_{\mathbb{R}^3} \frac{|\psi(x',z')|^2}{\sqrt{(x-x')^2 + \varepsilon^2|z-z'|^2}} dx' dz' \\ &= \int_{\mathbb{R}^2} \int_{|x-x'|<1} \frac{|\psi(x',z')|^2 - |\psi(x,z')|^2}{\sqrt{(x-x')^2 + \varepsilon^2|z-z'|^2}} dx' dz' \\ &\quad + \int_{\mathbb{R}^2} \int_{|x-x'|<1} \frac{|\psi(x,z')|^2}{\sqrt{(x-x')^2 + \varepsilon^2|z-z'|^2}} dx' dz' \\ &\quad + \int_{\mathbb{R}^2} \int_{|x-x'|\ge 1} \frac{|\psi(x',z')|^2}{\sqrt{(x-x')^2 + \varepsilon^2|z-z'|^2}} dx' dz' \\ &= I_1 + I_2 + I_3 \, . \end{aligned} \tag{2.5}$$

We first analyze the term I_1 by rewriting it as

$$I_1 = \int_{\mathbb{R}^2} \int_{|x-x'|<1} \frac{|\psi(x',z')|^2 - |\psi(x,z')|^2}{|x-x'|} dx' dz' + r_1^\varepsilon, \tag{2.6}$$

where r_1^ε is to be upper-bounded later. Using

$$|\psi(x,z) - \psi(x',z)| \le C|x-x'|^{1/2} \left(\int_{\mathbb{R}} |\partial_x \psi(y,z)|^2 \, dy \right)^{1/2}, \tag{2.7}$$

we deduce that the first term on the right-hand side is well defined and can be bounded thanks to (2.4):

$$\begin{aligned} \left| \int_{\mathbb{R}^2} \int_{|x-x'|<1} \frac{|\psi(x',z')|^2 - |\psi(x,z')|^2}{|x-x'|} dx' dz' \right| &\le C \, \|\partial_x \psi\|_{L^2} \, \|\psi\|_{L^\infty_x L^2_z} \int_0^1 \frac{1}{\xi^{1/2}} d\xi \\ &\le C \|\psi\|_{\mathcal{H}}^2. \end{aligned}$$

In order to estimate the remainder r_1^ε, we remark that for all $\gamma \in [0,2]$ there holds

$$0 \leq \frac{1}{|x-x'|} - \frac{1}{\sqrt{(x-x')^2+\varepsilon^2|z-z'|^2}} \leq \frac{\varepsilon^\gamma |z-z'|^\gamma}{|x-x'|^{1+\gamma}}. \tag{2.8}$$

Pick β such that $0<\beta<\min(\frac{1}{2},\frac{\alpha}{2})$ and take $\gamma=\beta$. One can estimate the remainder as

$$|r_1^\varepsilon| \leq C\varepsilon^\beta \int_{\mathbb{R}^2}\int_{|x-x'|<1} \frac{\|\partial_x\psi(\cdot,z')\|_{L^2_x}}{|x-x'|^{1/2+\beta}} \left(|z|^\beta+|z'|^\beta\right)\left(|\psi(x',z')|+|\psi(x,z')|\right) dx'dz',$$

where we used (2.8) and (2.7). By the Cauchy–Schwarz estimate, for all $u\in\mathcal{H}$ we get

$$\begin{aligned}\|r_1^\varepsilon u\|_{L^2} \leq C\varepsilon^\beta \|\partial_x\psi\|_{L^2}\left(\||z|^\beta\psi\|_{L^2}\,\|u\|_{L^\infty_x L^2_z} + \|\psi\|_{L^\infty_x L^2_z}\,\||z|^\beta u\|_{L^2}\right)\int_0^1 \frac{1}{\xi^{1/2+\beta}}d\xi \\ + C\varepsilon^\beta\|\partial_x\psi\|_{L^2}\|uw\|_{L^2}\end{aligned}$$

with

$$w(x) := \int_{|x-x'|<1} \frac{(\int_{\mathbb{R}^2}|z'|^{2\beta}|\psi(x',z')|^2dz')^{1/2}}{|x-x'|^{1/2+\beta}}dx'\,.$$

The first line of the right-hand side is bounded thanks to (2.4) and $\beta<\frac{1}{2}$. To bound the last term, we use Hölder and Hardy–Littlewood–Sobolev inequalities:

$$\|uw\|_{L^2} \leq \|w\|_{L^{1/\beta}}\,\|u\|_{L_x^{2/(1-2\beta)}L^2_z} \leq C\,\||z|^\beta\psi\|_{L^2}\,\|u\|_{L_x^{2/(1-2\beta)}L^2_z} \leq C\,\|\psi\|_{\mathcal{H}}\,\|u\|_{\mathcal{H}}\,,$$

where we used (2.4) and the fact that $\beta<\frac{\alpha}{2}$ and $\frac{2}{1-2\beta}>2$. Finally, we have

$$\|r_1^\varepsilon u\|_{L^2} \leq C\,\varepsilon^\beta\,\|\psi\|^2_{\mathcal{H}}\,\|u\|_{\mathcal{H}}.$$

For the term I_2, a direct computation of the integral with respect to x' gives

$$I_2 = 2(-\log\varepsilon+\log 2)\langle|\psi(x,\cdot)|^2\rangle - 2\int_{\mathbb{R}^2}\log|z-z'||\psi(x,z')|^2dz' + r_2^\varepsilon\,, \tag{2.9}$$

with

$$r_2^\varepsilon = 2\int_{\mathbb{R}^2}|\psi(x,z')|^2\log\left(\frac{1+\sqrt{1+\varepsilon^2|z-z'|^2}}{2}\right)dz'.$$

Let us first estimate the dominant term in (2.9). The term $\langle|\psi(x,\cdot)|^2\rangle$ is clearly bounded in L^∞ by (2.4). In order to bound the second term

$$v = \int_{\mathbb{R}^2}\log|z-z'||\psi(x,z')|^2dz',$$

we remark that

$$|\log|z-z'|| \leq C\left(\frac{\mathbf{1}_{|z-z'|<1}}{|z-z'|^{1/2}} + 1 + |z|^{\alpha/2} + |z'|^{\alpha/2}\right),$$

and from Hardy–Littlewood–Sobolev and Gagliardo–Nirenberg inequalities we get, pointwise in x,

$$\int_{|z-z'|<1} \frac{1}{|z-z'|^{1/2}} |\psi(x,z')|^2 dz' \leq C\|\psi(x,\cdot)\|_{L^4}^2 \leq C\|\psi(x,\cdot)\|_{L^2}\|\partial_z\psi(x,\cdot)\|_{L^2}.$$

Hence, for all $u \in \mathcal{H}$,

$$\begin{aligned}\|uv\|_{L^2} \leq C\,\|\psi\|_{L^\infty_x L^2_z}\,\|\partial_z\psi\|_{L^2}\,\|u\|_{L^\infty_x L^2_z} + C\||z|^{\alpha/2}\psi\|_{L^2}^2\,\|u\|_{L^\infty_x L^2_z} \\ + C\|\psi\|_{L^\infty_x L^2_z}^2\,\|(1+|z|^{\alpha/2})u\|_{L^2} \leq C\,\|\psi\|_{\mathcal{H}}^2\,\|u\|_{\mathcal{H}},\end{aligned}$$

where we used (2.4). Let us now estimate the remainder r_2^ε. With the above choice of $\beta \leq \frac{1}{2} < 2$, we have

$$\log\left(\frac{1+\sqrt{1+t^2}}{2}\right) \leq Ct^\beta \qquad \forall t > 0,$$

and thus, for all $u \in \mathcal{H}$,

$$\begin{aligned}\|r_2^\varepsilon u\|_{L^2} &\leq C\,\varepsilon^\beta\|\psi\|_{L^\infty_x L^2_z}(\||z|^\beta\psi\|_{L^2}\,\|u\|_{L^\infty_x L^2_z} + \|\psi\|_{L^\infty_x L^2_z}\,\||z|^\beta u\|_{L^2}) \\ &\leq C\,\varepsilon^\beta\,\|\psi\|_{\mathcal{H}}^2\,\|u\|_{\mathcal{H}},\end{aligned}$$

where we used again (2.4) and $\beta < \frac{\alpha}{2}$.

Consider now the term I_3, which we write as

$$I_3 = \int_{|x-x'|\geq 1} \frac{\left\langle |\psi(x',z')|^2 \right\rangle}{|x-x'|} dx' + r_3^\varepsilon, \tag{2.10}$$

with the following immediate bound for the dominant term:

$$0 \leq \int_{|x-x'|\geq 1} \frac{\left\langle |\psi(x',z')|^2 \right\rangle}{|x-x'|} dx' \leq \|\psi\|_{L^2}^2\,.$$

Moreover, from (2.8), the following estimate can be deduced for the remainder:

$$\begin{aligned}|r_3^\varepsilon| &\leq C\,\varepsilon^\beta \int_{\mathbb{R}^2}\int_{|x-x'|\geq 1} \frac{|z-z'|^\beta\,|\psi(x',z')|^2}{|x-x'|^{1+\beta}} dx'dz' \\ &\leq C\,\varepsilon^\beta\,\left(|z|^\beta\|\psi\|_{L^2}^2 + \||z|^{\beta/2}\,\psi\|_{L^2}^2\right).\end{aligned}$$

This is enough to conclude that

$$\|r_3^\varepsilon\,u\|_{L^2} \leq C\,\varepsilon^\beta\,\|\psi\|_{\mathcal{H}}^2\,\|u\|_{\mathcal{H}}\,.$$

To complete the proof of the proposition, it suffices to gather (2.6), (2.9), and (2.10), and then to use (2.1). ▯

3. Proof of the main theorem. As we said in the introduction, the system (1.5)–(1.7) admits two natural conservation laws: the conservation of the L^2 norm (1.16) and the energy estimate (1.17). Whereas it is immediate to deduce from the first one a uniform estimate of the L^2 norm of ψ^ε, let us examine the second one. Multiplied by ε^2, it gives

$$\|H_z^{1/2}\psi^\varepsilon(t)\|^2_{L^2(\mathbb{R}^3)} \leq \varepsilon^2\|\partial_x\psi_0^\varepsilon\|^2_{L^2(\mathbb{R}^3)} + \|\psi_0^\varepsilon\|^2_{\mathcal{H}} + \frac{\varepsilon^2}{|\log\varepsilon|}\|V^\varepsilon(0)|\psi_0^\varepsilon|^2\|_{L^1(\mathbb{R}^3)}\,. \tag{3.1}$$

From Proposition 2.1 and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}\frac{1}{|\log\varepsilon|}\|V^\varepsilon(0)|\psi_0^\varepsilon|^2\|_{L^1} \leq\ & \frac{1}{|\log\varepsilon|}\|V^\varepsilon(0)\psi_0^\varepsilon\|_{L^2}\|\psi_0^\varepsilon\|_{L^2}\\ &= \frac{1}{4\pi|\log\varepsilon|}\|G^\varepsilon(\psi_0)\,\psi_0^\varepsilon\|_{L^2}\|\psi_0^\varepsilon\|_{L^2} \leq C\,\|\psi_0^\varepsilon\|^4_{\mathcal{H}},\end{aligned} \tag{3.2}$$

and thus

$$\|H_z^{1/2}\psi^\varepsilon(t)\|^2_{L^2(\mathbb{R}^3)} \leq \|\psi_0^\varepsilon\|^2_{\mathcal{H}} + C\,\varepsilon^2\|\psi_0^\varepsilon\|^4_{\mathcal{H}}, \tag{3.3}$$

which is uniformly bounded, thanks to Assumption 1.2. In order to have a bound for ψ^ε in the energy space $\mathcal{H}$, it remains to bound the L^2 norm of $\partial_x\psi^\varepsilon$. This is done in the next subsection for well-prepared initial data.

3.1. Energy estimate for well-prepared data. We name "well-prepared data" a sequence of initial data $(\psi_0^\varepsilon)_{\varepsilon>0}$ in $\mathcal{H}$ which are polarized on the first eigenmode χ_1 of H_z, associated with the eigenvalue E_1 in the sense (1.18).

We now prove that, under Assumptions 1.1 and 1.2 and the assumption of well-prepared data, estimate (1.19) holds true. This relies only on the two conservation laws (1.16) and (1.17). Since E_1 is the bottom of the spectrum of H_z, we have

$$\|H_z^{1/2}u\|^2_{L^2} - E_1\|u\|^2_{L^2} = \int_{\mathbb{R}^3}\overline{u}\,(H_z - E_1)u\,dxdz = \|(H_z-E_1)^{1/2}u\|^2_{L^2}\,;$$

thus subtracting $\frac{E_1}{\varepsilon^2}\times$ (1.16) from (1.17) leads to the identity

$$\begin{aligned}\|\partial_x\psi^\varepsilon(t)\|^2_{L^2} + \frac{1}{\varepsilon^2}\|(H_z-E_1)^{1/2}\psi^\varepsilon(t)\|^2_{L^2} + \frac{1}{|\log\varepsilon|}\left\|V^\varepsilon(t)|\psi^\varepsilon(t)|^2\right\|_{L^1}&\\ = \|\partial_x\psi_0^\varepsilon\|^2_{L^2} + \frac{1}{\varepsilon^2}\|(H_z-E_1)^{1/2}\psi_0^\varepsilon\|^2_{L^2} + \frac{1}{|\log\varepsilon|}\left\|V^\varepsilon(0)|\psi_0^\varepsilon|^2\right\|_{L^1}&.\end{aligned}$$

By Assumption 1.2, (1.18), and (3.2), the right-hand side of this inequality is bounded independently of ε. Hence

$$\|\partial_x\psi^\varepsilon(t)\|^2_{L^2} + \frac{1}{\varepsilon^2}\|(H_z-E_1)^{1/2}\psi^\varepsilon(t)\|^2_{L^2} \leq C. \tag{3.4}$$

This estimate has two consequences. First, with (3.3) it gives

$$\|\psi^\varepsilon(t)\|_{\mathcal{H}} \leq C, \tag{3.5}$$

uniformly with respect to t. Second, this estimate shows that ψ^ε remains polarized on the first mode for all time. More precisely, denote

$$r^\varepsilon(t,x,z) = \psi^\varepsilon(t,x,z) - \chi_1(z)\int\psi^\varepsilon(t,x,z')\chi_1(z')dz'.$$

Remarking that $(H_z - E_1)^{1/2} \geq (E_2 - E_1)^{1/2} > 0$ in the operator sense on $\mathcal{H}$, the following estimate can be deduced from (3.4):

$$\|r^\varepsilon(t)\|_{\mathcal{H}} \leq C\varepsilon. \tag{3.6}$$

3.2. The convergence theorem. In this section, we prove the convergence result stated in Theorem 1.3. Let

$$\psi^\varepsilon(t,x,z) = e^{-itE_1/\varepsilon^2}\varphi_1^\varepsilon(t,x)\chi_1(z) + r^\varepsilon(t,x,z). \tag{3.7}$$

Inserting (3.7) into (1.5) and projecting on $Span(\chi_1)$ leads to the following equation:

$$i\partial_t \varphi_1^\varepsilon = -\partial_x^2 \varphi_1^\varepsilon + \frac{e^{itE_1/\varepsilon^2}}{|\log\varepsilon|}\int_{\mathbb{R}^2} V^\varepsilon(t,x,z)\psi^\varepsilon(t,x,z)\chi_1(z)dz. \tag{3.8}$$

To deal with the nonlinear term, we use the decomposition given by Proposition 2.1, with $V^\varepsilon = \frac{1}{4\pi}G^\varepsilon(\psi^\varepsilon)$. Remarking that, by orthogonality, we have

$$\left\langle |\psi^\varepsilon|^2 \right\rangle = |\varphi_1^\varepsilon|^2 + \left\langle |r^\varepsilon|^2 \right\rangle,$$

we get from (2.2)

$$\frac{e^{itE_1/\varepsilon^2}}{|\log\varepsilon|}\int_{\mathbb{R}^2} V^\varepsilon(t,x,z)\psi^\varepsilon(t,x,z)\chi_1(z)dz = \frac{1}{2\pi}|\varphi_1^\varepsilon|^2\varphi_1^\varepsilon + f^\varepsilon,$$

with

$$f^\varepsilon = \frac{1}{2\pi}\left\langle |r^\varepsilon|^2 \right\rangle \varphi_1^\varepsilon + \frac{e^{itE_1/\varepsilon^2}}{4\pi|\log\varepsilon|}\int_{\mathbb{R}^2}\left(R_1(\psi^\varepsilon) + R_2^\varepsilon(\psi^\varepsilon)\right)\psi^\varepsilon\chi_1 dz.$$

We clearly have

$$\|f^\varepsilon\|_{L^2(\mathbb{R})} \leq C\left\|\left\langle |r^\varepsilon|^2\right\rangle\right\|_{L^\infty(\mathbb{R})}\|\psi^\varepsilon\|_{L^2} + \frac{C}{|\log\varepsilon|}\left(\|R_1(\psi^\varepsilon)\psi^\varepsilon\|_{L^2} + \|R_2^\varepsilon(\psi^\varepsilon)\psi^\varepsilon\|_{L^2}\right).$$

In order to bound the first term, we notice that by the Cauchy–Schwarz estimate,

$$\left|\partial_x \left\langle |r^\varepsilon|^2\right\rangle\right|^{1/2} = \frac{\left|\mathcal{R}e \int \overline{r^\varepsilon}\,\partial_x r^\varepsilon\,dz\right|}{\left\langle |r^\varepsilon|^2\right\rangle^{1/2}} \leq \|\partial_x r^\varepsilon\|_{L^2_z},$$

and thus by the Sobolev embedding $H^1(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R})$,

$$\left\|\left\langle |r^\varepsilon|^2\right\rangle\right\|_{L^\infty(\mathbb{R})} \leq C\left\|\left\langle |r^\varepsilon|^2\right\rangle^{1/2}\right\|_{H^1(\mathbb{R})}^2 \leq C\left(\|r^\varepsilon\|_{L^2(\mathbb{R}^3)}^2 + \|\partial_x r^\varepsilon\|_{L^2(\mathbb{R}^3)}^2\right) \leq C\varepsilon^2,$$

where we used (3.6). Therefore, one deduces directly from (2.3) and (3.5) that

$$\|f^\varepsilon\|_{L^2(\mathbb{R})} \leq \frac{C}{|\log\varepsilon|}. \tag{3.9}$$

Now, the conclusion stems from a stability result for the cubic NLS equation in dimension one. Indeed, the functions φ_1^ε and φ solve, respectively,

$$i\partial_t\varphi_1^\varepsilon = -\partial_x^2\varphi_1^\varepsilon + \frac{1}{2\pi}|\varphi_1^\varepsilon|^2\varphi_1^\varepsilon + f^\varepsilon\,, \qquad \varphi_1^\varepsilon(0,x) = \int_{\mathbb{R}}\psi_0^\varepsilon(x,z)\chi_1(z)dz \tag{3.10}$$

and

$$i\partial_t \varphi = -\partial_x^2 \varphi + \frac{1}{2\pi}|\varphi|^2\varphi\,, \qquad \varphi(0,x) = \int_{\mathbb{R}} \psi_0(x,z)\chi_1(z)dz. \tag{3.11}$$

We remark that both functions are bounded in $H^1(\mathbb{R})$, and thus in $L^\infty(\mathbb{R})$, uniformly in time. For φ_1^ε this property is a direct consequence of (3.5), as $\|\varphi_1^\varepsilon\|_{H^1(\mathbb{R})} \leq \|\psi^\varepsilon\|_{\mathcal{H}}$. For φ, this stems from the energy conservation for the defocusing NLS equation (3.11) and from the fact that, by Assumption 1.2, the initial data $\varphi(0,\cdot)$ belongs to $H^1(\mathbb{R})$. Then it is easily seen that for all t we have

$$\begin{aligned}\|\varphi_1^\varepsilon(t) - \varphi(t)\|_{L^2} &\leq \|\psi_0^\varepsilon - \psi_0\|_{L^2} + \int_0^t \left(\frac{1}{2\pi}\left\||\varphi_1^\varepsilon|^2\varphi_1^\varepsilon - |\varphi|^2\varphi\right\|_{L^2} + \|f^\varepsilon(s)\|_{L^2}\right) ds \\ &\leq \|\psi_0^\varepsilon - \psi_0\|_{L^2} + C\int_0^t \|\varphi_1^\varepsilon(s) - \varphi(s)\|_{L^2} ds + \int_0^t \|f^\varepsilon(s)\|_{L^2} ds;\end{aligned}$$

so it follows from (3.9), from Assumption 1.2, and from the Gronwall lemma that for all $T > 0$

$$\|\varphi_1^\varepsilon - \varphi\|_{L^\infty([-T,T],L^2(\mathbb{R}))} \xrightarrow[\varepsilon\to 0]{} 0.$$

3.3. Towards a more precise approximation. According to (3.9), the convergence rate in Theorem 1.3 is at most $\mathcal{O}\left(\frac{1}{|\log\varepsilon|}\right)$. To go further, Proposition 2.1 suggests the form of the next term in the approximation of the initial model. Taking into account the R^1 term, one can consider the following system:

$$i\partial_t \widetilde{\varphi} = -\partial_x^2 \widetilde{\varphi} + \frac{1}{2\pi}|\widetilde{\varphi}|^2\widetilde{\varphi} + \frac{1}{4\pi|\log\varepsilon|}\left(\gamma|\widetilde{\varphi}|^2 + \mathrm{FP}\int_{\mathbb{R}} \frac{|\widetilde{\varphi}(x')|^2}{|x-x'|}\,dx'\right)\widetilde{\varphi}, \tag{3.12}$$

where

$$\gamma = -\int_{\mathbb{R}^4} \log\left(\frac{|z-z'|^2}{4}\right)|\chi_1(z)|^2|\chi_1(z')|^2 dz dz'$$

and

$$\widetilde{\varphi}(0,x) = \int_{\mathbb{R}^2} \psi_0^\varepsilon(x,z)\chi_1(z)dz.$$

From the approximation result given by Proposition 2.1, one could expect a better convergence rate:

$$\|\psi^\varepsilon(t,x,z) - e^{itH_z/\varepsilon^2}\widetilde{\varphi}(t,x)\chi_1(z)\|_{L^2} \leq C\,\varepsilon^\beta$$

with $\beta > 0$ as in Proposition 2.1. At the level of this article, this refined convergence result is a conjecture, as is the existence of the solution $\widetilde{\varphi}$ of (3.12). These questions will be investigated in a future work.

REFERENCES

[AFS] T. Ando, B. Fowler, and F. Stern, *Electronic properties of two-dimensional systems*, Rev. Mod. Phys., 54 (1982), pp. 437–672.

[Bas] G. Bastard, *Wave Mechanics Applied to Semi-conductor Heterostructures*, Les Éditions de Physique, EDP Sciences, Les Ulis Cedex, France, 1992.

[BCM] N. Ben Abdallah, F. Castella, and F. Méhats, *Time averaging for the strongly confined nonlinear Schrödinger equation, using almost periodicity*, J. Differential Equations, 245 (2008), pp. 154–200.

[BAMP] N. Ben Abdallah, F. Méhats, and O. Pinaud, *Adiabatic approximation of the Schrödinger–Poisson system with a partial confinement*, SIAM J. Math. Anal., 36 (2005), pp. 986–1013.

[BMSW] N. Ben Abdallah, F. Méhats, C. Schmeiser, and R. M. Weishäupl, *The nonlinear Schrödinger equation with strong anisotropic harmonic potential*, SIAM J. Math. Anal., 37 (2005), pp. 189–199.

[BM] F. Brezzi and P. A. Markowich, *The three dimensional Wigner–Poisson problem: Existence, uniqueness and approximation*, Math. Methods Appl. Sci., 14 (1991), pp. 35–61.

[Cas] F. Castella, *L^2 solutions to the Schrödinger-Poisson system: Existence, uniqueness, time behaviour, and smoothing effects*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1051–1083.

[Caz] T. Cazenave, *Semilinear Schrödinger Equations*, Lecture Notes AMS, AMS, New York, 2003.

[CDR] H. Cornean, P. Duclos, and B. Ricaud, *Effective models for excitons in carbon nanotubes*, Ann. Henri Poincaré, 8 (2007), pp. 135–163.

[FG] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press, Cambridge, UK, 1997.

[IZL] R. Illner, P. F. Zweifel, and H. Lange, *Global existence, uniqueness and asymptotic behaviour of solutions of the Wigner-Poisson and Schrödinger-Poisson systems*, Math. Methods Appl. Sci., 17 (1994), pp. 349–376.

[RS] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 1–4, Academic Press, New York, San Francisco, London, 1972–1979.

[VW] B. Vinter and C. Weisbuch, *Quantum Semiconductor Structures: Fundamentals & Applications*, Academic Press, New York, 1991.

SIAM J. Appl. Math.
Vol. 69, No. 4, pp. 1174–1204

BLOOMING IN A NONLOCAL, COUPLED PHYTOPLANKTON-NUTRIENT MODEL*

A. ZAGARIS[†], A. DOELMAN[†], N. N. PHAM THI[‡], and B. P. SOMMEIJER[§]

Abstract. Recently, it has been discovered that the dynamics of phytoplankton concentrations in an ocean exhibit a rich variety of patterns, ranging from trivial states to oscillating and even chaotic behavior [J. Huisman, N. N. Pham Thi, D. M. Karl, and B. P. Sommeijer, *Nature*, 439 (2006), pp. 322–325]. This paper is a first step towards understanding the bifurcational structure associated with nonlocal coupled phytoplankton-nutrient models as studied in that paper. Its main subject is the linear stability analysis that governs the occurrence of the first nontrivial stationary patterns, the *deep chlorophyll maxima* (DCMs) and the *benthic layers* (BLs). Since the model can be scaled into a system with a natural singularly perturbed nature, and since the associated eigenvalue problem decouples into a problem of Sturm–Liouville type, it is possible to obtain explicit (and rigorous) bounds on, and accurate approximations of, the eigenvalues. The analysis yields bifurcation-manifolds in parameter space, of which the existence, position, and nature are confirmed by numerical simulations. Moreover, it follows from the simulations and the results on the eigenvalue problem that the asymptotic linear analysis may also serve as a foundation for the secondary bifurcations, such as the oscillating DCMs, exhibited by the model.

Key words. phytoplankton, singular perturbations, eigenvalue analysis, Sturm–Liouville, Airy functions, WKB

AMS subject classifications. 35B20, 35B32, 34B24, 34E20, 86A05, 92D40

DOI. 10.1137/070693692

1. Introduction.

Phytoplankton forms the foundation of most aquatic ecosystems [16]. Since it transports significant amounts of atmospheric carbon dioxide into the deep oceans, it may play a crucial role in climate dynamics [6]. Therefore, the dynamics of phytoplankton concentrations have been studied intensely and from various points of view (see, for instance, [7, 11, 15] and the references therein). Especially relevant and interesting patterns exhibited by phytoplankton are the *deep chlorophyll maxima* (DCMs), or *phytoplankton blooms*, in which the phytoplankton concentration exhibits a maximum at a certain, well-defined depth of the ocean (or, in general, of a vertical water column). Simple, one-dimensional, scalar—but nonlocal—models for the influence of a depth-dependent light intensity on phytoplankton blooms have been studied since the early 1980s [14]. The nonlocality of these models is a consequence of the influence of the accumulated plankton concentration on the light intensity at a certain depth z (see (1.2) below). Numerical simulations and various mathematical approaches (see [5, 7, 8, 10, 12]) show that these models may, indeed, exhibit DCMs, depending on the manner in which the decay of the light intensity with depth is modeled and for certain parameter combinations.

*Received by the editors June 5, 2007; accepted for publication (in revised form) October 6, 2008; published electronically January 30, 2009. This work was supported by the Netherlands Organisation for Scientific Research (NWO).

http://www.siam.org/journals/siap/69-4/69369.html

[†]Korteweg-de Vries Institute, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands, and Centrum Wiskunde & Informatica (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (A.Zagaris@cwi.nl, A.Doelman@cwi.nl).

[‡]ABN AMRO Bank N.V., P.O. Box 283, 1000 EA, Amsterdam, The Netherlands (Nga.Pham.Thi@nl.abnamro.com).

[§]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (B.P.Sommeijer@cwi.nl).

The analysis in [14] establishes that, for a certain (large) class of light intensity functions, the scalar model has a stationary global attractor. This attractor may be trivial; i.e., the phytoplankton concentration W may decrease with time to $W \equiv 0$. If this trivial pattern is spectrally unstable, either the global attractor is a DCM or the phytoplankton concentration is maximal at the surface of the ocean (this latter case is called a *surface layer* (SL) [10, 15]). It should be noted here that *benthic layers* (BLs) [15]—i.e., phytoplankton blooms that become maximum at the bottom of the water column—cannot occur in the setting of [14], due to the choice of boundary conditions. Although the analysis in [14] cannot be applied directly to all scalar models in the literature, the main conclusion—that such models may only exhibit stationary nontrivial patterns (DCMs, SLs, or BLs)—seems to be true for each one of these models.

In sharp contrast to this, it has been numerically discovered recently [11] that systems—i.e., nonscalar models in which the phytoplankton concentration W is coupled to an evolution equation for a nutrient N—may exhibit complex behavior ranging from periodically oscillating DCMs to chaotic dynamics. These nonstationary DCMs have also been observed in the Pacific Ocean [11].

In this paper, we take a first step towards understanding the rich dynamics of the phytoplankton-nutrient models considered in [11]. Following [11], we consider the one-dimensional (i.e., depth-dependent only), nonlocal model,

$$\left\{ \begin{array}{l} W_t = D\,W_{zz} - V\,W_z + [\mu\,P(L,N) - l]\,W, \\ N_t = D\,N_{zz} - \alpha\,\mu\,P(L,N)\,W, \end{array} \right. \tag{1.1}$$

for $(z,t) \in [0, z_B] \times \mathbf{R}_+$ and where $z_B > 0$ determines the depth of the water column. The system is assumed to be in the turbulent mixing regime (see, for instance, [5, 10]), and thus the diffusion coefficient D is taken to be identically the same for W and N. The parameters V, l, α, and μ measure, respectively, the sinking speed of phytoplankton, the species-specific loss rate, the conversion factor, and the maximum specific production rate, and they are all assumed to be positive (see Remark 1.1 also). The light intensity L is modeled by

$$L(z,t) = L_I\,\mathrm{e}^{-K_{bg}z - R\int_0^z W(\zeta,t)\,d\zeta}, \tag{1.2}$$

where L_I is the intensity of the incident light at the water surface, K_{bg} is the light absorption coefficient due to nonplankton components, and R is the light absorption coefficient due to the plankton. Note that L is responsible for the introduction of nonlocality into the system. The function $P(L,N)$, which is responsible for the coupling, models the influence of light and nutrient on the phytoplankton growth, and it is taken to be

$$P(L,N) = \frac{LN}{(L+L_H)(N+N_H)}, \tag{1.3}$$

where L_H and N_H are the half-saturation constants of light and nutrient, respectively. We note that, from a qualitative standpoint, the particular form of P is of little importance. Different choices for P yield the same qualitative results, as long as they share certain common characteristics with the function given in (1.3); see Remark 1.1. Finally, we equip the system with the boundary conditions

$$D\,W_z - V\,W|_{z=0,z_B} = 0, \quad N_z|_{z=0} = 0, \quad \text{and} \quad N|_{z=z_B} = N_B, \tag{1.4}$$

i.e., no-flux through the boundaries except at the bottom of the column where N is at its maximum (prescribed by N_B). We refer the reader to Remark 1.1 for a discussion of more general models. To recast the model in nondimensional variables, we rescale time and space by setting

$$x = z/z_B \in (0,1) \quad \text{and} \quad \tau = \mu t \geq 0;$$

we introduce the scaled phytoplankton concentration ω, nutrient concentration η, and light intensity j,

$$\omega(x,\tau) = \frac{l\alpha z_B^2}{DN_B}\, W(z,t), \quad \eta(x,\tau) = \frac{N(z,t)}{N_B}, \quad j(x,\tau) = \frac{L(z,t)}{L_I};$$

and thus we recast (1.1) in the form

$$\text{(1.5)} \qquad \begin{cases} \omega_\tau = \varepsilon\omega_{xx} - \sqrt{\varepsilon}a\,\omega_x + (p(j,\eta) - \ell)\omega, \\ \eta_\tau = \varepsilon\left(\eta_{xx} - \frac{1}{\ell}p(j,\eta)\omega\right). \end{cases}$$

Here,
(1.6)

$$j(x,\tau) = \exp\left(-\kappa x - r\int_0^x \omega(s,\tau)\,ds\right), \quad \text{with} \quad \kappa = K_{bg}z_B \quad \text{and} \quad r = \frac{RDN_B}{l\alpha z_B},$$

and

$$\text{(1.7)} \qquad \varepsilon = \frac{D}{\mu z_B^2}, \quad a = \frac{V}{\sqrt{\mu D}}, \quad \ell = \frac{l}{\mu}, \quad \text{and} \quad p(j,\eta) = \frac{j\eta}{(j+j_H)(\eta+\eta_H)},$$

where $j_H = L_H/L_I$, $\eta_H = N_H/N_B$. The rescaled boundary conditions are given by

$$\text{(1.8)} \qquad \left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(0) = \left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(1) = 0, \quad \eta_x(0) = 0, \quad \text{and} \quad \eta(1) = 1.$$

These scalings are suggested by realistic parameter values in the original model (1.1) as reported in [11]. Typically,

$$D \approx 0.1\,\text{cm}^2/\text{s}, \quad V \approx 4.2\,\text{cm/h}, \quad z_B \approx 3\cdot 10^4\,\text{cm}, \quad l \approx 0.01/\text{h}, \quad \text{and} \quad \mu \approx 0.04/\text{h},$$

so that

$$\text{(1.9)} \qquad \varepsilon \approx 10^{-5}, \quad a \approx 1, \quad \text{and} \quad \ell \approx 0.25$$

in (1.5). Thus, realistic choices of the parameters in (1.1) induce a *natural singularly perturbed structure* in the model, as is made explicit by the scaling of (1.1) into (1.5). In this article, ε will be considered as an asymptotically small parameter, i.e., $0 < \varepsilon \ll 1$.

The simulations in [11] indicate that the DCMs bifurcate from the trivial stationary pattern,

$$\text{(1.10)} \qquad \bar{\omega}(x,\tau) \equiv 0, \quad \bar{\eta}(x,\tau) \equiv 1 \quad \text{for all } (x,\tau) \in [0,1] \times \mathbf{R}_+;$$

see also section 3. To analyze this (first) bifurcation, we set

$$(\omega(x,\tau), \eta(x,\tau)) = \left(\tilde{\omega}e^{\lambda\tau}, 1 + \tilde{\eta}e^{\lambda\tau}\right), \quad \text{with} \quad \lambda \in \mathbf{C},$$

and consider the (spectral) stability of $(\bar{\omega}, \bar{\eta})$. This yields the linear eigenvalue problem

$$\left\{ \begin{array}{r} \varepsilon\omega_{xx} - \sqrt{\varepsilon}a\,\omega_x + (f(x) - \ell)\omega = \lambda\omega, \\ \varepsilon\left(\eta_{xx} - \frac{1}{\ell}f(x)\omega\right) = \lambda\eta, \end{array} \right. \tag{1.11}$$

where we have dropped the tildes with a slight abuse of notation. The boundary conditions are

$$\left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(0) = \left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(1) = 0 \quad \text{and} \quad \eta_x(0) = \eta(1) = 0, \tag{1.12}$$

while the function f is the linearization of the function $p(j, \eta)$,

$$f(x) = \frac{1}{(1 + \eta_H)(1 + j_H \mathrm{e}^{\kappa x})}. \tag{1.13}$$

The linearized system (1.11) is *partially decoupled*, so that the stability of $(\bar{\omega}, \bar{\eta})$ as solution of the *two-component system* (1.5) is determined by two *one-component* Sturm–Liouville problems,

$$\begin{array}{c} \varepsilon\,\omega_{xx} - \sqrt{\varepsilon}\,a\,\omega_x + (f(x) - \ell)\omega = \lambda\omega, \\ \left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(0) = \left(\sqrt{\varepsilon}\omega_x - a\,\omega\right)(1) = 0, \end{array} \tag{1.14}$$

with η determined from the second equation in (1.11), and

$$\varepsilon\,\eta_{xx} = \lambda\eta \quad \text{with} \quad \eta_x(0) = \eta(1) = 0, \tag{1.15}$$

with ω identically equal to zero. The second of these problems, (1.15), is exactly solvable and describes the diffusive behavior of the nutrient in the absence of phytoplankton. Thus, it is not directly linked to the phytoplankton bifurcation problem that we consider, and we will not discuss it further. The phytoplankton behavior that we focus on is described by (1.14) instead, and hence we have returned to a scalar system as studied in [5, 7, 8, 10, 12, 14, 15]. However, our viewpoint differs significantly from that of those studies. The simulations in [11] (and section 3 of the present article) suggest that the destabilization of $(\bar{\omega}, \bar{\eta})$ into a DCM is merely the first in a series of bifurcations. In fact, section 3 shows that this DCM undergoes "almost immediately" a second bifurcation of Hopf type; i.e., it begins to oscillate periodically in time. According to [14], this is impossible in a scalar model (also, it has not been numerically observed in such models), and so the Hopf bifurcation must be induced by the *weak* coupling between ω and η in the full model (1.5).

Our analysis establishes that the largest eigenvalue λ_0 of (1.14) which induces the (stationary) DCM as it crosses through zero is the first of a sequence of eigenvalues λ_n that are only $\mathcal{O}(\varepsilon^{1/3})$ apart (see Figure 3.3, where $\varepsilon^{1/3} \approx 0.045$). The simulations in section 3 show that the distance between this bifurcation and the subsequent Hopf bifurcation of the DCM is of the same magnitude; see Figure 3.3 especially. Thus, the stationary DCM already destabilizes while λ_0 is still asymptotically small in ε, which indicates that the amplitude of the bifurcating DCM is also still asymptotically small and determined (at leading order) by $\omega_0(x)$, the eigenfunction associated with λ_0. This agrees fully with our linear stability analysis, since $\omega_0(x)$ indeed has the structure of a DCM (see sections 2 and 7). As a consequence, the *leading order* (in ε) stability analysis of the DCM is also governed by the partially decoupled system (1.11). In other words, although what drives the secondary bifurcation(s) is the coupling between $\omega(x)$ and $\eta(x)$ in (1.5), the leading order analysis is governed by the eigenvalues and eigenfunctions of (1.14). Naturally, the next eigenvalues and their associated eigenfunctions will play a key role in such a secondary bifurcation analysis, as will the eigenvalues and eigenfunctions of the trivial system (1.15).

Therefore, a detailed knowledge of the nature of the eigenvalues and eigenfunctions of (1.14) forms the foundation of analytical insight in the bifurcations exhibited by (1.5). This is the topic of the present paper; the subsequent (weakly) nonlinear analysis is the subject of work in progress.

The structure of the eigenvalue problem (1.14) is rather subtle, and therefore we employ two different analytical approaches. In sections 4–6, we derive explicit and rigorous bounds on the eigenvalues in terms of expressions based on the zeroes of the Airy function of the first kind and its derivative; see Theorem 2.1. We supplement this analysis with a WKB approach in section 7, where we show that the critical eigenfunctions have the structures of a DCM or a BL. This analysis establishes the existence of, first, the aforementioned sequence of eigenvalues that are $\mathcal{O}(\varepsilon^{1/3})$ apart, which is associated with the bifurcation of a DCM; and second, of another eigenvalue which also appears for biologically relevant parameter combinations and which is associated with the bifurcation of a BL—this bifurcation was not observed in [11]. This eigenvalue is isolated, in the sense that it is not part of the eigenvalue sequence associated with the DCMs—instead, it corresponds to a zero of a linear combination of the Airy function of the second kind and its derivative. Depending on the value of the dimensionless parameter a, the trivial state $(\bar{\omega}, \bar{\eta})$ bifurcates either into a DCM or into a BL. Our analysis establishes the bifurcation sets explicitly in terms of the parameters in the problem (section 2.2) and is confirmed by numerical simulations (section 3). Note that the codimension 2 point, at which DCM- and BL-patterns bifurcate simultaneously and which we determine explicitly, is related to that studied in [20]. Nevertheless, the differences are crucial—for instance, [20] considers a two-layer ODE model where, additionally, the DCM interacts with an SL instead of a BL (an SL cannot occur in our setting because $V > 0$ in (1.1); see Remark 1.1).

The outcome of our analysis is summarized in section 2, in which we also summarize the bio-mathematical interpretations of this analysis. We test and challenge the results of the stability analysis by numerical simulations of the full model in section 3. Although our insights are based only on linear predictions, and we do not yet have analytical results on the (nonlinear) stability of the patterns that bifurcate, we do find that there is an excellent agreement between the linear analysis and the numerical simulations. Thus, our analysis of (1.14) yields explicit bifurcation curves in the biological parameter space associated with (1.1). For any given values of the parameters, our analysis predicts whether one may expect a phytoplankton pattern with the structure of a (possibly oscillating) DCM, a pattern with the structure of a BL, or whether the phytoplankton will become extinct. Moreover, we also briefly consider secondary bifurcations into time-periodic patterns. These bifurcations are not directly covered by our linear analysis, but the distance between the first and second bifurcation in parameter space implies that the linearized system (1.14) must play a crucial role in the subsequent (weakly) nonlinear analysis; see the discussion above.

Remark 1.1. Our approach and findings for the model (1.1) (equivalently, (1.5)) are also applicable and relevant for more extensive models:

• In [11], (1.1) was extended to a model for various phytoplankton species $W_i(z,t)$ $(i = 1, \ldots, n)$. A stability analysis of the trivial pattern $W_i \equiv 0$, $N \equiv N_B$ yields n uncoupled copies of (1.14) in which the parameters depend on the species, i.e., on the index i. As a consequence, the results of this paper can also be applied to this multispecies setting.

- It is natural to include the possibility of horizontal flow and diffusion in the model (1.1). In the most simple setting, this can be done by allowing W and N to vary with (x, y, z, t) and to include horizontal diffusion terms in (1.1), i.e., $D_H(W_{xx}+W_{yy})$ and $D_H(N_{xx}+N_{yy})$ with $D_H \neq D$, in general—see [17], for instance. Again, the linear stability analysis of the trivial state is essentially not influenced by this extension. The exponentials in the ansatz following (1.10) now need to be replaced by $\exp(\lambda\tau + i(k_x\tilde{x} + k_y\tilde{y}))$, where k_x and k_y are wave numbers in the (rescaled) x and y directions. As a consequence, one only has to replace ℓ by $\ell - D_H(k_x^2 + k_y^2)$ in (1.14).
- The fact that we assign specific formulas to the growth and light intensity functions $P(L, N)$ (see (1.3)) and $L(z, t)$ (see (1.2)) is inessential for our analysis. One needs only that $f(x)$ is *decreasing* and *bounded* in $[0, 1]$—both assumptions are natural from a biological standpoint.
- We have considered "sinking" phytoplankton species in our model, i.e., $V > 0$ in (1.1) and thus $a > 0$ in (1.14). Our analysis can also be applied to buoyant species ($V \leq 0$). In that case, the bifurcating DCMs may transform into SLs—see also [10, 15].
- The values of ε, a, and ℓ in (1.9) are typical of oceanic settings [11]. These values differ in an estuary, and ε can no longer be assumed to be asymptotically small; see [19] and the references therein. Moreover, phytoplankton blooms in an estuary are strongly influenced by the concentration of suspended sediment and typically occur not only at a certain depth z, but also at a certain horizontal position in the estuary. Thus, (1.14) must be extended to account for such blooms; however, it may still play an important role as a limiting case or a benchmark [19].

2. The main results.

In the first part of this section, we present our main results in full mathematical detail. In section 2.2, we present a bio-mathematical interpretation of these results.

2.1. Mathematical analysis.

We define the parameter $\nu = 1/(1+\eta_H)$, the function F through

$$F(x) = F(x; j_H, \kappa, \nu) = f(0) - f(x) \geq 0 \quad \text{for all} \quad x \in [0,1] \tag{2.1}$$

(see (1.13)), and the constants $\sigma_L = \sigma_L(\kappa, j_H, \nu)$ and $\sigma_U = \sigma_U(\kappa, j_H, \nu)$ so that

$$\sigma_L\, x \leq F(x) \leq \sigma_U\, x \quad \text{for all} \quad x \in [0,1]. \tag{2.2}$$

The optimal values of σ_U and σ_L can be determined explicitly. This (simple yet technical) analysis is postponed until after the formulation of Theorem 2.1; see Lemma 2.1 and Figures 2.2 and 2.3. Next, we define the parameters

$$A = \frac{a^2}{4}, \quad \beta = \sqrt{\frac{A}{\sigma}}, \quad \text{and} \quad 0 < \gamma \equiv \left(\frac{\varepsilon}{\sigma}\right)^{1/3} \ll 1, \tag{2.3}$$

with a as in (1.7) and σ an a priori parameter. (Later, σ will be set equal to either σ_L or σ_U.) Furthermore, we write Ai and Bi for the Airy functions of the first and second kind [1], respectively, and $A_n < 0$, $n \in \mathbf{N}$, for the nth zero of $\mathrm{Ai}(x)$; see Figure 2.1. We also define the functions

$$\Gamma(\mathrm{Ai}, x) = \mathrm{Ai}(x) - \sqrt{\gamma}\,\beta^{-1}\,\mathrm{Ai}'(x) \quad \text{and} \quad \Gamma(\mathrm{Bi}, x) = \mathrm{Bi}(x) - \sqrt{\gamma}\,\beta^{-1}\,\mathrm{Bi}'(x) \tag{2.4}$$

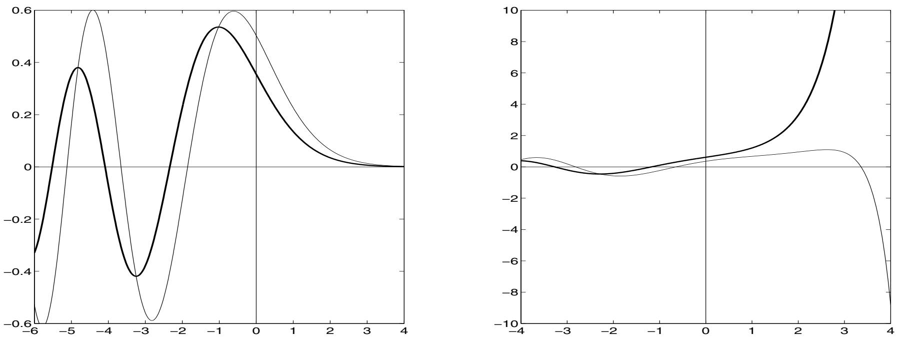


FIG. 2.1. *Left: Airy function of the first kind (thick line) plotted with the function* $\Gamma\,(\mathrm{Ai},\cdot)$ *(thin line). Right: Airy function of the second kind (thick line) plotted with* $\Gamma\,(\mathrm{Bi},\cdot)$ *(thin line). Here,* $\varepsilon = 0.1$*,* $a = 3$*, and* $\sigma = 2$*.*

(see Figure 2.1 and section 5.1) and write $A'_{n,\sigma}$ for the nth zero of $\Gamma\,(\mathrm{Ai}, x)$ $(n \in \mathbf{N})$—which is $\mathcal{O}(\sqrt{\gamma})$ close to A_n—and $B_{0,\sigma}$ for the positive zero of $\Gamma\left(\mathrm{Bi}, \gamma^{-1}(1+x)\right)$—which exists for all $\beta > 1$ and is equal to $\beta^2 - 1$ at leading order in γ; see Lemma A.2 for more accurate estimates. Finally, we let

$$\lambda^* = f(0) - \ell - A, \quad \lambda_0^{*,\sigma} = \lambda^* + A\,\beta^{-2}\,B_{0,\sigma}, \quad \lambda_n^{*,\sigma} = \lambda^* - \gamma\,A\,\beta^{-2}\,\left|A'_{n,\sigma}\right|, \quad n \in \mathbf{N}, \tag{2.5}$$

and we note that $\lambda_0^{*,\sigma}$ and $\lambda_n^{*,\sigma}$ are decreasing functions of σ. We can now formulate our main result.

THEOREM 2.1. *Let* $M \in \mathbf{N}$*. There exists an* $\varepsilon_0 > 0$ *and a constant* $C > 0$ *such that, for all* $0 < \varepsilon < \varepsilon_0$ *and* $0 \le n \le M$*, the first* $M+1$ *eigenvalues* $\lambda_0 > \cdots > \lambda_M$ *of* (1.14) *satisfy the following:*

(a) *For each* $0 < \sigma_U < A$*, there exists a constant* $B > 0$ *such that*

$$\lambda_0^{*,\sigma_U} - C\,\varepsilon^{2/3}\,\mathrm{e}^{-B/\sqrt{\varepsilon}} \le \lambda_0 \le \lambda_0^{*,\sigma_L} + C\,\varepsilon^{2/3}\,\mathrm{e}^{-B/\sqrt{\varepsilon}}$$

and

$$\lambda_n^{*,\sigma_U} - C\,\varepsilon^{1/6}\,\mathrm{e}^{-B/\sqrt{\varepsilon}} \le \lambda_n \le \lambda_n^{*,\sigma_L} + C\,\varepsilon^{1/6}\,\mathrm{e}^{-B/\sqrt{\varepsilon}} \quad \textit{for all} \quad 1 \le n \le M.$$

(b) *For each* $\sigma_L > A$*, there exists a constant* $B > 0$ *such that*

$$\lambda_{n+1}^{*,\sigma_U} - C\,\varepsilon^{1/6}\,\mathrm{e}^{-B/\sqrt{\varepsilon}} \le \lambda_n \le \lambda_{n+1}^{*,\sigma_L} + C\,\varepsilon^{1/6}\,\mathrm{e}^{-B/\sqrt{\varepsilon}} \quad \textit{for all} \quad 0 \le n \le M.$$

Theorem 2.1 and (2.5) establish that, for any $M \in \mathbf{N}$ and for sufficiently small $\varepsilon > 0$ (equivalently, for sufficiently small $\gamma > 0$), all first $M+1$ eigenvalues of (1.14) are $\mathcal{O}(\varepsilon^{1/3})$ close to λ^*, except for the special eigenvalue λ_0 if $\sigma_U < A$. Both types of eigenvalues correspond to biologically relevant patterns in (1.1)—to DCMs and BLs, respectively; see section 2.2. This dependence on the parameters is quite subtle; further, the weakly nonlinear stability analysis must be based on a detailed understanding of the linear eigenvalue problem including all of the eigenmodes associated with the asymptotically close eigenvalues (see also the introduction). As a result, the required analysis becomes rather extensive. For this reason, we defer the proof of Theorem 2.1 to sections 4–6.

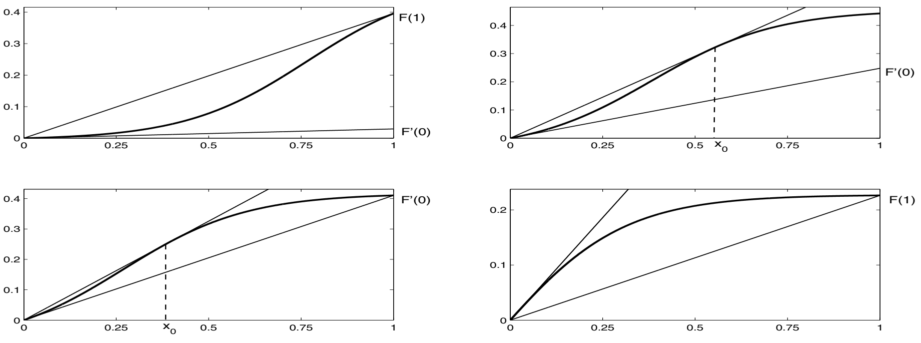


FIG. 2.2. *The function F (thick curve) and the linear functions bounding it (thin lines). Here, $\eta_H = 1$, $\kappa = 6$, and $j_H = 0.01 < j_H^{(1)}$ (top left panel), $j_H^{(1)} < j_H = 0.1 < j_H^{(2)}$ (top right panel), $j_H^{(2)} < j_H = 0.2 < 1$ (bottom left panel), and $j_H = 1.2 > 1$ (bottom right panel).*

Moreover, this analysis establishes that the bounds on the eigenvalues are, up to exponentially small terms, explicitly given in terms of zeroes of the Airy functions $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$ (and their derivatives (2.4)) and of the bounds $\sigma_L\, x$ and $\sigma_U\, x$ on $F(x)$ in (2.2). This enables us (by unscaling) to explicitly quantify the regions in the parameter space associated with (1.1) in which DCMs or BLs can be expected to appear (see sections 2.2 and 3).

The following lemma provides explicit control on $\sigma_L\, x$ and $\sigma_U\, x$.

LEMMA 2.1. *Let*

$$j_H^{(1)}(\kappa) = \frac{\mathrm{e}^{-\kappa} - 1 + \kappa}{\mathrm{e}^{\kappa} - 1 - \kappa} \qquad \textit{and} \qquad j_H^{(2)}(\kappa) = \frac{\mathrm{e}^{-\kappa}}{j_H^{(1)}(\kappa)},$$

so that $0 < j_H^{(1)}(\kappa) < j_H^{(2)}(\kappa) < 1$ for all $\kappa > 0$. Also, for all $\kappa > 0$ and $j_H \in (j_H^{(1)}(\kappa), 1)$, define the point $x_0 = x_0(\kappa, j_H) \in (0,1)$ via $F(x_0) = x_0\, F'(x_0)$. Then,

$$(2.6) \qquad \sigma_L = \begin{cases} F'(0), & 0 < j_H \le j_H^{(2)}, \\ F(1), & j_H > j_H^{(2)}, \end{cases} \qquad \sigma_U = \begin{cases} F(1), & 0 < j_H \le j_H^{(1)}, \\ F'(x_0), & j_H^{(1)} < j_H < 1, \\ F'(0), & j_H \ge 1, \end{cases}$$

and

$$(2.7) \qquad \sigma_L(\kappa, j_H, \nu) = \nu\, \sigma_L(\kappa, j_H, 1), \qquad \sigma_U(\kappa, j_H, \nu) = \nu\, \sigma_U(\kappa, j_H, 1).$$

This lemma is proved by straightforward calculus. Figures 2.2 and 2.3 give a graphical representation of the lemma for various representative subcases.

As we shall see in section 3, the eigenvalue bounds established in Theorem 2.1 are quite sharp and predict very well the bifurcations of the full unscaled model (1.1). Nevertheless, the rigorous analysis of sections 4–6 yields no information on the characteristics of the associated eigenfunctions, which are of particular interest to the nature of the patterns generated by (1.1) as λ_0 crosses through zero (see section 3). Moreover, the width of the intervals bounding the eigenvalues of (1.14) is of the same order in ε—namely of $\mathcal{O}(\varepsilon^{1/3})$—as the distance between successive eigenvalues. This

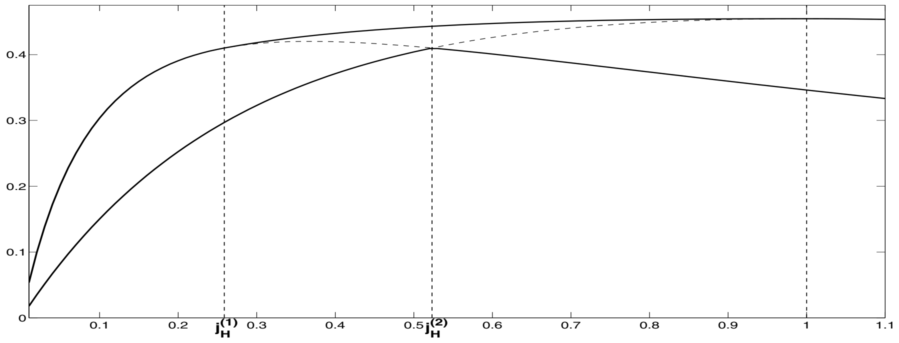


FIG. 2.3. *The quantities σ_U (upper thick curve), σ_L (lower thick curve), $F(1)$ (dashed curve to the left), and $F'(0)$ (dashed curve to the right) as functions of j_H and for $\eta_H = 0.1$, $\kappa = 2$. Note that $F(1)$ merges with σ_U for $j_H \le j_H^{(1)}$ and with σ_L for $j_H \ge j_H^{(2)}$, while $F'(0)$ merges with σ_L for $j_H \le j_H^{(2)}$ and with σ_U for $j_H \ge 1$. Also note that the WKB method (see section 7) yields that the location of the eigenvalue close to $\lambda_0^{*,\sigma}$ (see Theorem 2.1) is determined by $F(1)$, at leading order, whereas the locations of the eigenvalues close to $\lambda_n^{*,\sigma}$, $n \in \mathbf{N}$, are determined by $F'(0)$ at leading order.*

is especially relevant in the transitional case $\sigma_L < A < \sigma_U$, for which Theorem 2.1 offers no information.

For these reasons, we complete our analysis of (1.14) with an asymptotic WKB approximation (section 7). We derive asymptotic formulas for the eigenvalues and for the corresponding eigenfunctions. Using these formulas, we show the following:

- In case (a) of Theorem 2.1, the profile of the eigenfunction ω_0 corresponding to the largest eigenvalue λ_0 is of boundary layer type near the bottom. In terms of the phytoplankton concentration, this profile corresponds to a BL.
- In case (b) of the same theorem, ω_0 has the shape of a spike around the point $x = x_{\mathrm{DCM}}$, where x_{DCM} is determined, to leading order in ε, by $F(x_{\mathrm{DCM}}) = A$ (see Figure 7.1). This profile corresponds to a DCM around x_{DCM}.
- The transitional region between cases (a) and (b) in Theorem 2.1 is described, to leading order in ε, by the equation $A = F(1)$. Indeed, the leading order approximation to λ_0 is

(2.8)
$$\lambda_{0,0} = f(1) - \ell \qquad \text{in the region } F(1) = f(0) - f(1) < A \text{ (and } \omega_0 \text{ is a BL)},$$
(2.9)
$$\lambda_{0,0} = \lambda^* = f(0) - \ell - A \quad \text{in the region } F(1) = f(0) - f(1) > A \text{ (and } \omega_0 \text{ is a DCM)}.$$

Recalling Lemma 2.1, we see that this transition occurs at a value of A which is, to leading order in ε, equal to σ_U when $0 < j_H \le j_H^{(1)}$, equal to σ_L when $j_H \ge j_H^{(2)}$, and between σ_U and σ_L when $j_H^{(1)} < j_H < j_H^{(2)}$.

2.2. Bio-mathematical interpretation. The agreement between the numerical simulations and the field data reported in [11] establishes the biological relevance of model problem (1.1) and of its dynamics. This paper contains the first steps towards a bio-mathematical understanding of this model, especially in relation to the existing models in the literature that exhibit only simple, stationary patterns [5, 7, 8, 10, 12, 14].

The fact that (1.1) can be scaled into the singularly perturbed equation (1.5) for biologically relevant choices of the parameters is essential to the analysis in this paper. Moreover, together with the linear stability analysis, these scalings enable us to understand the fundamental structure of the twelve-dimensional parameter space associated with (1.1) and its boundary conditions (1.4) (in the biologically relevant region). In fact, it follows from Theorem 2.1 and (2.8)–(2.9) that the dimensionless parameters A, ℓ, $f(0)$, and $f(1)$, which are defined in section 2.1, are the main parameter combinations in the model as they capture its most relevant biological aspects.

Our stability analysis determines the regions in parameter space in which phytoplankton may persist, i.e., in which the trivial solution of (1.1) and (1.4) corresponding to absence of phytoplankton ($W(z,t) \equiv 0$ in (1.1)) is unstable. In that case, nontrivial patterns with $W(z,t) > 0$, for all t, bifurcate from the trivial solution, which implies that the model admits stable, positive phytoplankton populations. Theorem 2.1 establishes the existence of two distinct types of phytoplankton populations at onset. One is formed by a large—in fact infinite—family of "DCM-modes" and occurs for A below the threshold value $f(0) - f(1)$; the region where these modes become stable is determined by $\lambda^* = f(0) - \ell - A$; see (2.9). Within this family, the phytoplankton concentrations are negligible for most z, except for a certain localized (spatial) region in which the phytoplankton population is concentrated—see Figure 7.1 in which the first, most unstable member of this family is plotted (in scaled coordinates). These are the DCM-patterns observed in [11]. Our analysis shows that many different DCM-patterns appear almost instantaneously. More precisely, as a parameter enters into the region in which the trivial solution is unstable, a succession of asymptotically close bifurcations in which different types of DCM-patterns are created takes place. In other words, even asymptotically close to onset, there are many competing DCM-modes. This partly explains why the "pure" DCM-mode as represented in Figure 7.1 can be observed only very close to onset (see [11] and section 3.2): it may be destabilized by the competition with other modes.

The second type of phytoplankton population that may appear at onset occurs for A above the threshold value $f(0) - f(1)$ and has the structure of a BL: the phytoplankton population is concentrated near $z = z_B$, i.e., at the bottom of the water column. Unlike the DCM-modes, there is a *single* BL-mode; the region where this mode becomes stable is determined, in this case, by $f(1)-\ell$; see (2.8). This mode may also dominate the dynamics of (1.1) in a part of the biologically relevant parameter space, as may be seen in section 3.2. Note that the BL-mode has not been observed in [11]; naturally, this is hardly surprising since one can sample numerically only a very limited region of a twelve-dimensional parameter space. From the biological point of view, the fact that the model (1.1) allows for attractors of the BL type may be the most important finding of this paper. Like DCMs, BLs have been observed in field data (see [15] and references therein). The analysis here quantifies the parameter values for which DCM- or BL-patterns occur. Hence, our results may be used to determine oceanic regions and/or phytoplankton species for which BLs may be expected to exist. It would be even more interesting to locate a setting in which DCMs and BLs interact, as they are expected to do because of the existence of the codimension 2 point at which the (first) DCM-mode and the BL-mode bifurcate simultaneously; see section 3.

3. Bifurcations and simulations.

3.1. The bifurcation diagram. In this section, we use the WKB expressions (2.8)–(2.9) for the first few eigenvalues to identify the bifurcations that system (1.14) undergoes. In this way, we identify the regions in the parameter space where the BL

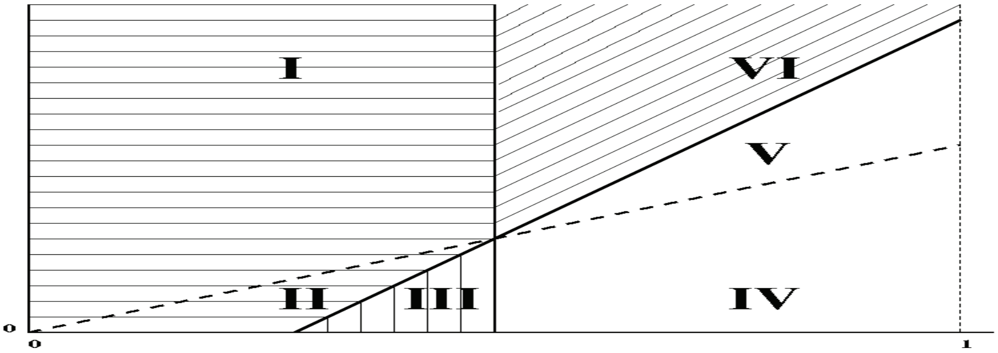


FIG. 3.1. *The bifurcation diagram in the (ν, A)-plane. The horizontal axis corresponds to $\nu = 1/(1+\eta_H)$, while the vertical one corresponds to $A = a^2/4$. In the region shaded horizontally, the trivial zero state is stable. In the region shaded vertically, DCMs bifurcate, while BL profiles remain damped. In the region shaded diagonally, BL profiles bifurcate, while DCM profiles remain damped. Finally, in the unshaded region, both profiles grow linearly.*

and DCM steady states become stable. As already mentioned in the Introduction, we are primarily interested in the effect of environmental conditions—in particular, of nutrient concentration and diffusion—on phytoplankton. For this reason, we choose to vary the parameters $\eta_H = N_H/N_B$ (which encapsulates information pertaining to the nutrient levels and nutrient absorption by phytoplankton) and $a = V/\sqrt{\mu D}$ (which is a measure of diffusion; see (1.7)). The remaining four dimensionless parameters (ε, κ, j_H, and ℓ) are kept constant. We recall here the definitions $\nu = 1/(1+\eta_H)$ and $A = a^2/4$.

The curves separating the regions in the (ν, A)-plane which are characterized by qualitatively different behavior of the rescaled model (1.5), (1.8) may be found by recasting (2.9) and (2.8) in terms of the rescaled parameters. In particular, using (1.13), (2.1), and (2.5), we find (see Figure 3.1) the following:

- In regions I and II, λ_0 is given, to leading order, by (2.8) (in region I) and by (2.9) (in region II). In either case, $\lambda_0 < 0$, and hence the zero (trivial) state is stable.
- In region III, λ_0 is given by (2.9) and is positive. In fact, the further into this region one goes, the more eigenvalues cross zero and become positive, since they are $\mathcal{O}(\varepsilon^{1/3})$ apart by Theorem 2.1. All of these eigenvalues are associated with DCMs.
- In region VI, λ_0 is given by (2.8) and is positive, while all other eigenvalues are negative. Thus, the only bifurcating patterns in this regime are BL profiles.
- Finally, in regions IV and V, eigenvalues associated with both BL and DCM profiles are positive, and thus no further info can be derived from our linear analysis.

The boundaries of these regions may be deduced explicitly in the aforementioned manner. First, setting the expression for λ_0 in (2.8) equal to zero, we obtain, to leading order, the vertical line separating the regions I, II, and III from the regions IV, V, and VI,

$$\nu = \ell\,(1 + \mathrm{e}^{\kappa} j_H).$$

Next, setting the expression for λ_0 in (2.9) equal to zero, we obtain, to leading order, the diagonal line separating the regions I, II, and VI from III, IV, and V,

$$A = \frac{1}{1+j_H}\,\nu - \ell.$$

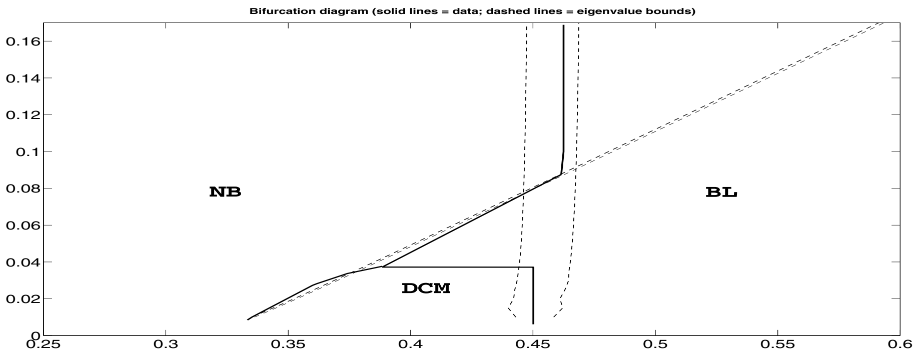


FIG. 3.2. *The bifurcation diagram in the* (ν, A)*-plane for* $\varepsilon = 9 \cdot 10^{-5}$, $\ell = 0.2$, $j_H = 0.5$, $\kappa = 1$. *("NB" stands for "no blooming.") The solid curves correspond to numerical simulations, while the dashed ones correspond to the bounds predicted theoretically; see Theorem* 2.1.

Finally, setting the expressions for λ_0 in (2.8) and (2.9) equal to each other, we obtain the transitional regime $A = F(1)$. In terms of the rescaled parameters, we find

$$A = \left(\frac{1}{1 + j_H} - \frac{1}{1 + \mathrm{e}^{\kappa} j_H} \right) \nu.$$

Since the physical region $n_H > 0$ corresponds to the region $0 < \nu < 1$, these formulas imply that

(a) for $0 < \ell < (1 + \mathrm{e}^{\kappa} j_H)^{-1}$, both a BL and a DCM may bifurcate,

(b) for $(1 + \mathrm{e}^{\kappa} j_H)^{-1} < \ell < (1 + j_H)^{-1}$, only a DCM may bifurcate,

(c) for $\ell > (1 + j_H)^{-1}$, the trivial state is stable.

Remark 3.1. Similar information may be derived by the rigorous bounds in Theorem 2.1, with the important difference that the dividing curves have to be replaced by regions of finite thickness.

3.2. Numerical simulations. In this section, we present numerical simulations on the full model (1.1)–(1.4), and we compare the results with our theoretical predictions. The parameters are chosen in biologically relevant regions [11].

We considered first the validity of our asymptotic analysis; i.e., we checked whether the analytically obtained bounds for the occurrence of the DCMs and BLs—see Theorem 2.1, section 3.1, Figure 3.1, and Remark 3.1—can be recovered by numerical simulations of the PDE (1.1)–(1.4). We used the numerical method described in Remark 3.2 at each node of a two-dimensional grid of a part of the (ν, A)-parameter plane (keeping all other parameters fixed) to determine the attracting pattern generated by (1.1)–(1.4) and chose the initial profile at each node in the parameter space to be the numerically converged pattern for an adjacent node at the previous step.

In Figure 3.2, we present the region near the codimension 2 point in the (ν, A)-parameter plane at which both the DCMs and the BLs bifurcate (with all other parameters fixed: $\varepsilon = 9 \cdot 10^{-5}$, $\ell = 0.2$, $j_H = 0.5$, $\kappa = 1$). Away from this codimension 2 point, the numerically determined bifurcation curves are clearly within the bounds given by Theorem 2.1 and thus confirm our analysis. Note that this suggests that the bifurcations have a supercritical nature—an observation that does not follow from our linear analysis. Near the codimension 2 point, a slight discrepancy between

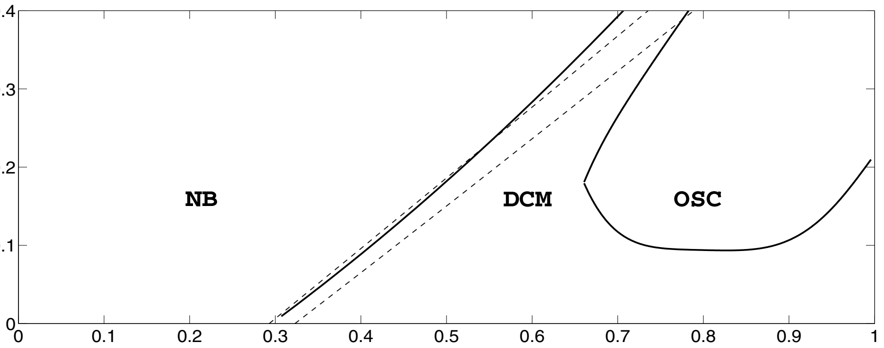


Fig. 3.3. *The bifurcation diagram in the* (ν, A)*-plane for* $\varepsilon = 9 \cdot 10^{-5}$, $\ell = 0.25$, $j_H = 0.033$, $\kappa = 20$. *Region NB corresponds to no blooming, and region OSC to oscillatory DCMs. The solid curves correspond to numerical simulations, and the dashed ones to the points at which* λ_0 *(left line) and* λ_1 *(right line) cross zero; see* (2.9) *and Figure* 3.1. *For these parameter values, the bifurcation of the BLs occurs in a nonphysical part of the domain.*

our analysis and the numerical findings becomes apparent. First, we note that the bifurcation from the trivial state (no phytoplankton) to the DCM state is not exactly in the region determined by Theorem 2.1. However, for this combination of parameters, this region is quite narrow—in fact, it is narrower than the width of the rectangular grid of the (ν, A)-parameter plane that we used to determine Figure 3.2, which implies that the simulations do not disagree with the analysis. The other discrepancy, namely the occurrence of a small "triangle" of BL patterns in the region where one would expect DCMs, is related to the presence of the codimension 2 point. To understand the true nature of the dynamics, one needs to perform a weakly nonlinear analysis near this point and, presumably, a more detailed numerical analysis that distinguishes between DCMs, BLs, and patterns that have the structure of a combined DCM and BL. This is the topic of work in progress.

Unlike the simulations presented in [11], here we considered the secondary bifurcations only briefly. Figure 3.3 shows the primary bifurcation of the trivial state into a DCM and the secondary bifurcation (of Hopf type) of the DCM into an oscillating DCM—see [11] for more (biological) details on this behavior. A priori, one would expect that our linear stability analysis of the trivial state could not cover this Hopf bifurcation. However, in Figure 3.3 we also plotted the leading order approximations of the curves at which the first two eigenvalues associated with the stability of the trivial state, λ_0 and λ_1, cross through the imaginary axis. It follows that the distance (in parameter space) between the primary and the secondary bifurcations is asymptotically small in ε, and similar to the distance between the successive eigenvalues λ_n. This observation is based on several simulations realized for different values of ε. It is crucial information for the subsequent (weakly) nonlinear analysis, since the fact that the DCM undergoes its secondary Hopf bifurcation for parameter combinations that are asymptotically close (in ε) to the primary bifurcation implies that the above a priori expectation is not correct; instead, the stability and bifurcation analysis of the DCM can, indeed, be based on the linear analysis presented here. The higher order eigenvalues $\lambda_1, \lambda_2, \ldots$, the associated eigenfunctions $\omega_1(x), \omega_2(x), \ldots$, and their "slaved" η-components $\eta_1(x), \eta_2(x), \ldots$ (which can be determined explicitly using (1.11)) will serve as necessary inputs for this nonlinear analysis.

Thus, a “full” linear stability analysis of the uncoupled system (1.14) as presented here may serve as a foundation for the analysis of secondary bifurcations that can only occur in the coupled system (see the introduction and [14]). This feature is very special and quite uncommon in explicit models. It is due to the *natural singularly perturbed nature* of the scaled system (1.5), and it provides an opportunity to obtain fundamental insight into phytoplankton dynamics. This analysis, including the aforementioned codimension 2 analysis and the associated secondary bifurcations of BLs, is the topic of work in progress.

Remark 3.2. The numerical results were obtained by the “Method of Lines” approach. First, we discretized the spatial derivatives approximating the diffusion terms in the model using second-order symmetric formulas and employing a third-order upwind-biased method to discretize the advection term (see [13] for the suitability of these schemes to the current problem). Next, we integrated the resulting system of ODEs forward in time with the widely used time-integration code VODE (see [3] and http://www.netlib.org/ode). Throughout all simulations, we combined a spatial grid of a sufficiently high resolution with a high precision time integration to ensure that the conclusions drawn from the simulations are essentially free of numerical errors.

4. Eigenvalue bounds.

As a first step towards the proof of Theorem 2.1, we recast (1.14) in a form more amenable to analysis. First, we observe that the operator involved in this eigenvalue problem is self-adjoint only if $a = 0$. Applying the Liouville transformation

$$w(x) = \mathrm{e}^{-\sqrt{A/\varepsilon}\,x}\omega(x) = \mathrm{e}^{-(\beta/\gamma^{3/2})\,x}\omega(x), \tag{4.1}$$

we obtain the self-adjoint problem

$$\begin{gathered} \varepsilon\, w_{xx} + (f(x) - \ell - A)w = \lambda\, w, \\ \left(\sqrt{\varepsilon}\, w_x - \sqrt{A}\, w\right)(0) = \left(\sqrt{\varepsilon}\, w_x - \sqrt{A}\, w\right)(1) = 0. \end{gathered}$$

Recalling (2.1) and (2.5), we write this equation in the form

$$\mathcal{L}\, w = \mu\, w, \quad \text{with} \quad \mathcal{G}\,(w,0) = \mathcal{G}\,(w,1) = 0. \tag{4.2}$$

The operator $\mathcal{L}$, the scalar μ, and the linear functionals $\mathcal{G}(\cdot, x)$ are defined by

$$\mathcal{L} = -\varepsilon\frac{d^2}{dx^2} + F(x), \quad \mu = \lambda^* - \lambda, \quad \mathcal{G}\,(w,x) = w(x) - \sqrt{\frac{\varepsilon}{A}}\, w_x(x). \tag{4.3}$$

This is the desired form of the eigenvalue problem (1.14). To prove Theorem 2.1, we decompose the operator $\mathcal{L}$ into a self-adjoint part for which the eigenvalue problem is exactly solvable and a positive definite part. Then, we use the following comparison principle to obtain the desired bounds.

THEOREM 4.1 (see [18, sections 8.12–8.13]). *Let $\hat{\mathcal{A}}$ and $\mathcal{A}$ be self-adjoint operators bounded below with compact inverses, and write their eigenvalues as $\hat{\mu}_0 \leq \hat{\mu}_1 \leq \cdots \leq \hat{\mu}_n \leq \cdots$ and $\mu_0 \leq \mu_1 \leq \cdots \leq \mu_n \leq \cdots$, respectively. If $\mathcal{A} - \hat{\mathcal{A}}$ is positive semidefinite, then $\hat{\mu}_n \leq \mu_n$ for all $n \in \{0, 1, \ldots\}$.*

4.1. Crude bounds for the eigenvalues of $\mathcal{L}$.

First, we derive crude bounds for the spectrum $\{\mu_n\}$ of $\mathcal{L}$ to demonstrate the method and establish that $\mathcal{L}$ satisfies the boundedness condition of Theorem 4.1.

LEMMA 4.1. *The eigenvalues μ_n satisfy the inequalities*

$$-A \leq \mu_0 \leq F(1) - A \quad \textit{and} \quad \varepsilon n^2\pi^2 \leq \mu_n \leq F(1) + \varepsilon n^2\pi^2, \quad n \in \mathbf{N}. \tag{4.4}$$

Proof. Let $c \in \mathbf{R}$. We start by decomposing $\mathcal{L}$ as

$$\mathcal{L} = \mathcal{L}^{0,c} + \mathcal{F}^{0,c}, \quad \text{where} \quad \mathcal{L}^{0,c} = -\varepsilon\frac{d^2}{dx^2} + c \quad \text{and} \quad \mathcal{F}^{0,c} = F(x) - c. \tag{4.5}$$

Then, we write $\{\mu_n^{0,c}\}$ for the set of eigenvalues of the problem

$$\mathcal{L}^{0,c}w^{0,c} = \mu^{0,c}w^{0,c}, \quad \text{with} \quad \mathcal{G}\left(w^{0,c}, 0\right) = \mathcal{G}\left(w^{0,c}, 1\right) = 0, \tag{4.6}$$

with the eigenvalues arranged so that $\mu_0^{0,c} \leq \mu_1^{0,c} \leq \cdots \leq \mu_n^{0,c} \leq \cdots$.

For $c = c_L = 0$, the operator $\mathcal{L}^{0,c_L}$ is self-adjoint, while $\mathcal{F}^{0,c_L} = F(x) \geq 0$ is a positive definite multiplicative operator. Thus, using Theorem 4.1, we obtain the inequalities

$$\mu_n^{0,c_L} \leq \mu_n \quad \text{for all} \quad n \in \mathbf{N} \cup \{0\}. \tag{4.7}$$

Next, for $c = c_U = F(1)$, the operator $\mathcal{F}^{0,c_U} = F(x) - F(1) \leq 0$ is negative definite, while $\mathcal{L}^{0,c_U}$ is self-adjoint. Hence, we may write

$$\mathcal{L}^{0,c_U} = \mathcal{L} - \mathcal{F}^{0,c_U},$$

where $-\mathcal{F}^{0,c_U}$ is now positive definite. The fact that the spectrum $\{\mu_n\}$ of $\mathcal{L}$ is bounded from below by (4.7) allows us to use Theorem 4.1 to bound each μ_n from above,

$$\mu_n \leq \mu_n^{0,c_U} \quad \text{for all} \quad n \in \mathbf{N} \cup \{0\}.$$

Combining this bound and (4.7), we obtain

$$\mu_n^{0,c_L} \leq \mu_n \leq \mu_n^{0,c_U} \quad \text{for all} \quad n \in \mathbf{N} \cup \{0\}. \tag{4.8}$$

Naturally, the eigenvalue problem (4.6) may be solved exactly to obtain

$$\mu_0^{0,c} = c - A \quad \text{and} \quad \mu_n^{0,c} = c + \varepsilon n^2\pi^2, \quad n \in \mathbf{N}. \tag{4.9}$$

Combining these formulas with (4.8), we obtain the inequalities (4.4). □

4.2. Tight bounds for the eigenvalues of $\mathcal{L}$. The accurate bounds for the eigenvalues of (4.2) described in Theorem 2.1 may be obtained by bounding F by linear functions; see (2.2) and Lemma 2.1. In the next lemma, we bound the eigenvalues μ_n by the eigenvalues $\mu_n^{1,\sigma}$ of a simpler problem. Then, in Lemma 4.3, we obtain strict, exponentially small bounds for $\mu_n^{1,\sigma}$.

LEMMA 4.2. *Let $\sigma \in \{\sigma_L, \sigma_U\}$, with σ_L and σ_U as defined in Lemma* 2.1*, define the operator $\mathcal{L}^{1,\sigma} = -\varepsilon\frac{d^2}{dx^2} + \sigma x$, and write $\{\mu_n^{1,\sigma}\}$ for the eigenvalues corresponding to the problem*

$$\mathcal{L}^{1,\sigma}w = \mu^{1,\sigma}w, \quad \textit{with} \quad \mathcal{G}(w,0) = \mathcal{G}(w,1) = 0. \tag{4.10}$$

Let $\{\mu_n^{1,\sigma}\}$ be arranged so that $\mu_0^{1,\sigma} \leq \mu_1^{1,\sigma} \leq \cdots \leq \mu_n^{1,\sigma} \leq \cdots$. Then,

$$\mu_n^{1,\sigma_L} \leq \mu_n \leq \mu_n^{1,\sigma_U} \quad \textit{for all} \quad n \in \mathbf{N} \cup \{0\}. \tag{4.11}$$

Proof. First, we decompose $\mathcal{L}$ as

$$\mathcal{L} = \mathcal{L}^{1,\sigma} + \mathcal{F}^{1,\sigma}, \quad \text{where} \quad \mathcal{L}^{1,\sigma} = -\varepsilon \frac{d^2}{dx^2} + \sigma x, \quad \mathcal{F}^{1,\sigma} = F(x) - \sigma x, \tag{4.12}$$

and $\sigma \in \{\sigma_L, \sigma_U\}$. We note here that $\mathcal{L}^{1,\sigma}$ is self-adjoint.

Next, $\mathcal{F}^{1,\sigma_L}$ is a positive definite multiplicative operator, since $F(x) \geq \sigma_L x$ (see (2.2)). Thus, $\mu_n^{1,\sigma_L} \leq \mu_n$ for all $n \in \mathbf{N} \cup \{0\}$, by Theorem 4.1. In contrast, $\mathcal{F}^{1,\sigma_U}$ is negative definite, since $F(x) \leq \sigma_U x$. Therefore, we write

$$\mathcal{L}^{1,\sigma_U} = \mathcal{L} - \mathcal{F}^{1,\sigma_U},$$

where now $-\mathcal{F}^{1,\sigma_U}$ is positive definite. The fact that the spectrum $\{\mu_n\}$ is bounded from below by Lemma 4.1 allows us to use Theorem 4.1 to bound each μ_n from above, $\mu_n \leq \mu_n^{1,\sigma_U}$. Combining both bounds for each n, we obtain (4.11). □

Hence, it remains to solve the eigenvalue problem (4.10). Although this problem is not explicitly solvable, the eigenvalues may be calculated up to terms exponentially small in ε. Letting

$$\mu_0^{*,\sigma} = \lambda^* - \lambda_0^{*,\sigma} = -A\,\beta^{-2}\,B_{0,\sigma} \quad \text{and} \quad \mu_n^{*,\sigma} = \lambda^* - \lambda_n^{*,\sigma} = \gamma\,A\,\beta^{-2}\,\left|A'_{n,\sigma}\right| > 0, \quad n \in \mathbf{N}, \tag{4.13}$$

where we have recalled the definitions in section 2, we can prove the following lemma.

Lemma 4.3. *Let $M \in \mathbf{N}$ be fixed, and define*

$$\delta_{0,\sigma} = \gamma^2 \exp\left(-\tfrac{2}{3}\gamma^{-3/2}\left[3(1+B_{0,\sigma}-B)^{3/2} - 2(B_{0,\sigma}-B)^{3/2} - (1+B_{0,\sigma}+B)^{3/2}\right]\right),$$
$$\delta_{n,\sigma} = \sqrt{\gamma}\,A^{1/6}\,\beta^{-1/3}\exp\left(-\tfrac{4}{3}\,\gamma^{-3/2} + 2\,|A_{n+1}|\,\gamma^{-1/2}\right) \quad \textit{for all} \quad 1 \leq n \leq M+1$$

and for all $0 < B < B_{0,\sigma}$ for which the exponent in the expression for $\delta_{0,\sigma}$ is negative. Then, for each such B, there exists an $\varepsilon_0 > 0$ and positive constants $C_0, \ldots, C_{M+1}$ such that, for all $0 < \varepsilon < \varepsilon_0$ and $0 \leq n \leq M$, the first $M+1$ eigenvalues $\mu_0^{1,\sigma}, \ldots, \mu_M^{1,\sigma}$ corresponding to (4.10) *satisfy the following:*

(a) *For $\beta > 1$, $\left|\mu_0^{1,\sigma} - \mu_0^{*,\sigma}\right| < C_0\,\delta_{0,\sigma}$ and $\left|\mu_n^{1,\sigma} - \mu_n^{*,\sigma}\right| < C_n\,\delta_{n,\sigma}$ for all $1 \leq n \leq M$.*

(b) *For $0 < \beta < 1$, $\left|\mu_n^{1,\sigma} - \mu_{n+1}^{*,\sigma}\right| < C_{n+1}\,\delta_{n+1,\sigma}$ for all $0 \leq n \leq M$.*

Lemmas 4.2 and 4.3 in combination with definitions (2.5) and (4.13) yield Theorem 2.1. The bounds on $\mu_0^{1,\sigma}, \ldots, \mu_M^{1,\sigma}$ are derived in section 5. The fact that these are indeed the $M+1$ first eigenvalues corresponding to (4.10) is proved in section 6. Note that Theorem 2.1 follows immediately from this lemma, in combination with the above analysis and the observation that the condition $\beta > 1$ is equivalent to $0 < \sigma < A$, and the condition $0 < \beta < 1$ equivalent to $\sigma > A$.

5. The eigenvalues $\boldsymbol{\mu_0^{1,\sigma}, \ldots, \mu_M^{1,\sigma}}$.

In this section, we derive the bounds on $\mu_0^{1,\sigma}, \ldots, \mu_M^{1,\sigma}$ of Lemma 4.3. In section 5.1, we reduce the eigenvalue problem (4.10) to the algebraic one of locating the roots of an Evans-type function $\mathcal{D}$. In section 5.2, we identify the roots of $\mathcal{D}$ with those of two functions $\mathcal{A}$ and $\mathcal{B}$ which are related to the Airy functions and simpler to analyze than $\mathcal{D}$. Finally, in section 5.3, we identify the relevant roots of $\mathcal{A}$ and $\mathcal{B}$ and thus also of $\mathcal{D}$.

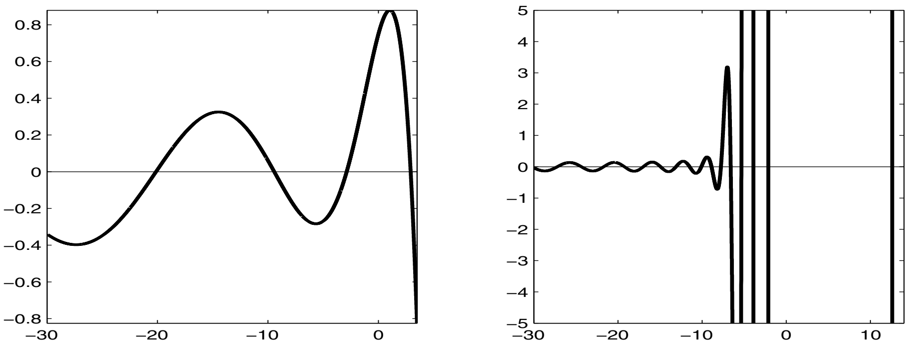


FIG. 5.1. *The function $\mathcal{D}(\bar{\chi})$ for $a = 3$, $\sigma = 1$, and $\varepsilon = 0.1$ (left panel), $\varepsilon = 0.001$ (right panel).*

5.1. Reformulation of the eigenvalue problem.

First, we derive an algebraic equation, the solutions of which correspond to the eigenvalues of (4.10). We start by rescaling the eigenvalue $\mu^{1,\sigma}$ and the independent variable x via

$$\bar{\chi} = -\gamma^{-1} A^{-1} \beta^2 \mu^{1,\sigma} \quad \text{and} \quad x = \gamma(\chi - \bar{\chi}). \tag{5.1}$$

Then, we define the linear functional

$$\Gamma(w, \bar{\chi}) = w(\bar{\chi}) - \sqrt{\gamma}\,\beta^{-1} w'(\bar{\chi}) \quad \text{for all differentiable functions } w, \tag{5.2}$$

and we remark that, for w equal to Ai or Bi, this definition agrees with that given in (2.4). Further introducing the Wronskian

$$\mathcal{D}(\bar{\chi}) = \Gamma(\mathrm{Ai}, \bar{\chi})\,\Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right) - \Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)\Gamma(\mathrm{Bi}, \bar{\chi}) \tag{5.3}$$

(see also Figure 5.1), we can prove the following lemma.

LEMMA 5.1. *The eigenvalue problem* (4.10) *has $\mu^{1,\sigma}$ as an eigenvalue if and only if $\mathcal{D}(\bar{\chi}) = 0$.*

Proof. Using (5.1), we rewrite problem (4.10) in the form

$$\begin{gathered} \frac{d^2 w}{d\chi^2} = \chi w, \quad \chi \in [\bar{\chi}, \gamma^{-1} + \bar{\chi}], \\ \Gamma(w, \bar{\chi}) = \Gamma\left(w, \gamma^{-1} + \bar{\chi}\right) = 0. \end{gathered} \tag{5.4}$$

This is an Airy equation and thus has the general solution

$$w(\chi) = D_A \,\mathrm{Ai}(\chi) + D_B \,\mathrm{Bi}(\chi). \tag{5.5}$$

The boundary conditions become

$$\begin{array}{rclcl} \Gamma(w, \bar{\chi}) & = & D_A \Gamma(\mathrm{Ai}, \bar{\chi}) + D_B \Gamma(\mathrm{Bi}, \bar{\chi}) & = & 0, \\ \Gamma\left(w, \gamma^{-1} + \bar{\chi}\right) & = & D_A \Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right) + D_B \Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right) & = & 0. \end{array} \tag{5.6}$$

The sufficient and necessary condition for the existence of nontrivial solutions to this system is that its determinant—which is the Wronskian $\mathcal{D}$ given in (5.3)—vanishes, and the lemma is proved. ☐

5.2. Product decomposition of the function $\mathcal{D}$. In the preceding section, we saw that the values of $\bar{\chi}$ corresponding to the eigenvalues $\mu^{1,\sigma}$ must be zeroes of $\mathcal{D}$. In the next section, we will prove that the first few zeroes of $\mathcal{D}$ are all $\mathcal{O}(1)$, in the case $0 < \beta < 1$, and both $\mathcal{O}(1)$ and $\mathcal{O}(\gamma^{-1})$ in the case $\beta > 1$. To identify them, we rewrite $\mathcal{D}$ in the form

$$\mathcal{D}(\bar{\chi}) = \Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right)\mathcal{A}(\bar{\chi}) = \Gamma\left(\mathrm{Ai}, \bar{\chi}\right)\mathcal{B}(\bar{\chi}), \tag{5.7}$$

where we have defined the functions

$$\mathcal{A}(\bar{\chi}) = \Gamma\left(\mathrm{Ai}, \bar{\chi}\right) - \frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)}{\Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right)}\Gamma\left(\mathrm{Bi}, \bar{\chi}\right), \tag{5.8}$$

$$\mathcal{B}(\bar{\chi}) = \Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right) - \frac{\Gamma\left(\mathrm{Bi}, \bar{\chi}\right)}{\Gamma\left(\mathrm{Ai}, \bar{\chi}\right)}\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right). \tag{5.9}$$

Here, $\mathcal{A}$ is well defined for all $\bar{\chi}$ such that $\Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right) \neq 0$, while $\mathcal{B}$ is well defined for all $\bar{\chi}$ such that $\Gamma\left(\mathrm{Ai}, \bar{\chi}\right) \neq 0$. Equation (5.7) implies that the roots of $\mathcal{A}$ and $\mathcal{B}$ are also roots of $\mathcal{D}$.

In the next section, we will establish that the $\mathcal{O}(1)$ roots of $\mathcal{D}$ coincide with roots of $\mathcal{A}$ and the $\mathcal{O}(\gamma^{-1})$ ones with roots of $\mathcal{B}$. To prove this, we first characterize the behaviors of $\mathcal{A}$ and $\mathcal{B}$ for $\mathcal{O}(1)$ and $\mathcal{O}(\gamma^{-1})$ values of $\bar{\chi}$, respectively, by means of the next two lemmas. In what follows, we write $E(x) = \exp(-(2/3)x^{3/2})$ for brevity and $||\cdot||_{[X_L,X_R]}$ for the W^1_∞-norm over any interval $[X_L, X_R]$,

$$||w||_{[X_L,X_R]} = \max_{\bar{\chi}\in[X_L,X_R]} |w(\bar{\chi})| + \max_{\bar{\chi}\in[X_L,X_R]} |w'(\bar{\chi})|. \tag{5.10}$$

LEMMA 5.2. *Let $X < 0$ be fixed. Then there is a $\gamma_0 > 0$ and a constant $c_A > 0$ such that*

$$||\mathcal{A}(\cdot) - \Gamma\left(\mathrm{Ai}, \cdot\right)||_{[X,0]} < c_A\, \gamma^{-1/2}\, E(\gamma^{-1}(2 + 3\,\gamma\, X)^{2/3}) \quad \text{for all} \quad 0 < \gamma < \gamma_0. \tag{5.11}$$

For the next lemma, we switch to the independent variable $\bar{\psi} = \gamma\bar{\chi}$ to facilitate calculations. We analyze the behavior of $\mathcal{B}(\gamma^{-1}\bar{\psi})$ for $\mathcal{O}(1)$ values of $\bar{\psi}$ (equivalently, for $\mathcal{O}(\gamma^{-1})$ values of $\bar{\chi}$) as $\gamma \downarrow 0$.

LEMMA 5.3. *Let $0 < \Psi_L < \Psi_R$ be fixed. Then there is a $\gamma_0 > 0$ and a constant $c_B > 0$ such that, for all $0 < \gamma < \gamma_0$,*

$$\begin{aligned} &\left|\left|E(\gamma^{-1}(1 + \bar{\psi}))\left[\mathcal{B}\left(\gamma^{-1}\bar{\psi}\right) - \Gamma\left(\mathrm{Bi}, \gamma^{-1}(1 + \bar{\psi})\right)\right]\right|\right|_{\bar{\psi}\in[\Psi_L,\Psi_R]} \\ &\qquad < c_B\, \gamma^{-1/4}\left[\frac{E(\gamma^{-1}(1 + \Psi_L))}{E(\gamma^{-1}\Psi_L)}\right]^2. \end{aligned}$$

The proofs of these lemmas are given in Appendices B and C, respectively.

5.3. Zeroes of $\mathcal{D}$. Using Lemma 5.2 and an auxiliary result, we can locate the roots of $\mathcal{D}$.

LEMMA 5.4. *Let $M \in \mathbf{N}$ be fixed, $A'_{n,\sigma}$ and $B_{0,\sigma}$ be defined as in section 2, and B, $\delta_{0,\sigma}, \ldots, \delta_{M,\sigma}$ be defined as in Lemma 4.3. Then, for each admissible B, there is a $\gamma_0 > 0$ and positive constants $c_0, \ldots, c_M$ such that, for all $0 < \gamma < \gamma_0$, $\mathcal{D}(\bar{\chi})$ has roots $\bar{\chi}_0 > \bar{\chi}_1 > \cdots > \bar{\chi}_M$ satisfying the following bounds:*

(a) *For $\beta > 1$,*

$$\left|\bar{\chi}_0 - \gamma^{-1}B_{0,\sigma}\right| < c_0\, \gamma^{-1}\, \delta_{0,\sigma} \quad \text{and} \quad \left|\bar{\chi}_n - A'_{n,\sigma}\right| < c_n\, \gamma^{-1}\, \delta_{n,\sigma} \quad \text{for all} \quad 1 \leq n \leq M.$$

(b) *For* $0 < \beta < 1$,

$$\left|\bar{\chi}_n - A'_{n+1,\sigma}\right| < c_n\, \gamma^{-1}\, \delta_{n+1,\sigma} \quad \textit{for all} \quad 0 \le n \le M.$$

The proof of this lemma requires the following elementary result.

LEMMA 5.5. *Let C and G be real-valued continuous functions and H be real-valued and differentiable. Let $\delta > 0$ and $z_0 \in [Z_L, Z_R] \subset \mathbf{R}$ be such that*

$$H(z_0) = 0, \quad \max_{[Z_L,Z_R]} H' = -H_0 < 0, \quad \max_{[Z_L,Z_R]} |C(G - H)| < \delta,$$
$$\textit{and} \quad \min_{[Z_L,Z_R]} C = C_0 > 0.$$

If $\delta < C_0 H_0 \min(z_0 - Z_L, Z_R - z_0)$, then G has a zero z_* such that $|z_* - z_0| \le \delta/(C_0 H_0)$.

Proof. Let $z_\ell = z_0 - \delta/(C_0 H_0)$ and $z_r = z_0 + \delta/(C_0 H_0)$. Since $Z_L < z_\ell < z_0 < z_r < Z_R$, we have

$$G(z_\ell) = H(z_\ell) + G(z_\ell) - H(z_\ell) \ge \int_{z_0}^{z_\ell} H'(z)\, dz - \frac{\max_{[Z_L,Z_R]} |C(G - H)|}{\min_{[Z_L,Z_R]} C}$$
$$> (z_0 - z_\ell) H_0 - \frac{\delta}{C_0} = 0.$$

Similarly, we may prove that $G(z_r) < 0$, and the desired result follows. □

Proof of Lemma 5.4. (a) First, we prove the existence of a root $\bar{\chi}_0$ satisfying the desired bound. We recall that $\bar{\psi}$ was defined above via $\bar{\psi} = \gamma\bar{\chi}$; hence, it suffices to show that there is a root $\bar{\psi}_0$ of $\mathcal{D}(\gamma^{-1}\bar{\psi})$ satisfying the bound $|\bar{\psi}_0 - B_{0,\sigma}| < c_0\, \delta_0$ for some $c_0 > 0$. Equation (5.7) reads $\mathcal{D}(\gamma^{-1}\bar{\psi}) = \Gamma\,(\mathrm{Ai}, \gamma^{-1}\bar{\psi})\, \mathcal{B}(\gamma^{-1}\bar{\psi})$. Here, $\Gamma\,(\mathrm{Ai}, \gamma^{-1}\bar{\psi})$ has no positive roots, by definition of Γ and because $\mathrm{Ai}(\gamma^{-1}\bar{\psi}) > 0$ and $\mathrm{Ai}'(\gamma^{-1}\bar{\psi}) < 0$ for all $\bar{\psi} > 0$. Thus, $\bar{\chi}_0$ must be a root of $\mathcal{B}$. Its existence and the bound on it follow from Lemmas 5.3 and 5.5. Indeed, let $z_0 = B_{0,\sigma}$, $Z_L = B_{0,\sigma} - B$, $Z_R = B_{0,\sigma} + B$, $C = E$ (see section 5.2), $G = B$, and $H = \Gamma\,(\mathrm{Bi}, \cdot)$. Lemma 5.3 provides a bound δ on $||C(G - H)||_{[Z_L,Z_R]}$. Also, using Corollary A.1, we may calculate

$$\begin{array}{rcccl} C_0 & = & \min_{[Z_L,Z_R]} E(\gamma^{-1}(1 + \bar{\psi})) & = & E(\gamma^{-1}(1 + Z_R)), \\ -H_0 & = & \max_{[Z_L,Z_R]} \Gamma\left(\mathrm{Bi}', \gamma^{-1}(1 + \bar{\psi})\right) & < & -c\,\gamma^{5/4}\left[E(\gamma^{-1}(1 + Z_L))\right]^{-1}. \end{array}$$

Now, δ satisfies the condition $\delta < C_0 H_0 B$ of Lemma 5.5 for all γ small enough. Thus, we may apply Lemma 5.5 to obtain the desired bound on $\bar{\chi}_0$.

Next, we show that $\mathcal{A}$ has the remaining roots $\bar{\chi}_1, \dots, \bar{\chi}_M$. We fix $A_{M+1} < X < A_M$ and let $I_1, \dots, I_M$ be disjoint intervals around the first M zeroes of Ai, $A_1, \dots, A_M$, respectively. Lemma 5.2 states that $\mathcal{A}(\bar{\chi})$ and $\Gamma\,(\mathrm{Ai}, \bar{\chi})$ are exponentially close in the W^1_∞-norm over $[X, 0]$. Thus, for all $0 < \gamma < \gamma_0$ (with γ_0 small enough), $\mathcal{A}$ has M distinct roots $\bar{\chi}_1 \in I_1, \dots, \bar{\chi}_M \in I_M$ in $[X, 0]$ by Lemma A.2. Since $\Gamma\left(\mathrm{Bi}, \gamma^{-1} + \bar{\chi}\right)$ can be bounded away from zero over $[X, 0]$ using Lemma A.1 (with $p = 1$ and $q = \bar{\chi}$), we conclude that $\mathcal{D}$ has the M distinct roots $\bar{\chi}_1, \dots, \bar{\chi}_M$ in $[X, 0]$.

(b) The argument used in part (a)—where $\beta > 1$—to establish the bounds on the $\mathcal{O}(1)$ roots of $\mathcal{A}$ does not depend on the sign of $\beta - 1$. Therefore, it applies also to this case—where $0 < \beta < 1$—albeit in an interval $[X, 0]$, with $A_{M+2} < X < A_{M+1}$, yielding $M + 1$ roots which we label $\bar{\chi}_0, \dots, \bar{\chi}_M$.

On the other hand, $B_{0,\sigma} < 0$ for $0 < \beta < 1$, because of the estimate on $B_{0,\sigma}$ in Lemma A.2. As a result, the argument used to identify that root does not apply any more, since now $B_{0,\sigma} < 0$ and thus Lemma 5.3 may not be applied to provide the bound δ needed in Lemma 5.5. In fact, were this root to persist and remain close to $\gamma^{-1}B_{0,\sigma}$ as in case (a), it would become *large* and *negative* by the estimate in Lemma A.2 and hence smaller than the roots $\bar{\chi}_0, \ldots, \bar{\chi}_M$ obtained above. Thus, it could never be the leading eigenvalue in this parameter regime. □

6. The eigenfunctions $w_0^{1,\sigma}, \ldots, w_M^{1,\sigma}$.

In the previous section, we located some of the eigenvalues $\mu^{1,\sigma}$. In this section, we show that the eigenvalues we identified are the largest ones. To achieve this, we derive formulas for the eigenfunctions $w_0^{1,\sigma}, \ldots, w_M^{1,\sigma}$ associated with $\mu_0^{1,\sigma}, \ldots, \mu_M^{1,\sigma}$, respectively, and show that $w_n^{1,\sigma}$ has n zeroes in the interval $[\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]$ (corresponding to the interval $[0,1]$ in terms of x; see (5.1)). The desired result follows, then, from standard Sturm–Liouville theory [4]. In particular, we prove the following lemma.

LEMMA 6.1. *Let $M \in \mathbf{N}$. Then, there is a $\gamma_0 > 0$ such that, for all $0 < \gamma < \gamma_0$ and for all $n = 0, 1, \ldots, M$, the eigenfunction $w_n^{1,\sigma}$ corresponding to the eigenvalue $\mu_n^{1,\sigma}$ has exactly n zeroes in the interval $[\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]$.*

The proof of this lemma occupies the rest of this section. Parallel to it, we show that the profile of ω_0 associated with w_0 through (4.1) is that of (a) a boundary layer near the bottom of the water column (BL) for $\beta > 1$, and (b) an interior, nonmonotone boundary layer (a *spike* [9]) close to the point $0 < x_{\mathrm{DCM}} = \beta^2 < 1$ (DCM) for $0 < \beta < 1$.

We start by fixing $\bar{\chi}$ to be $\bar{\chi}_n$, for some $n = 1, \ldots, M$. The corresponding eigenvalue is $\mu_n^{1,\sigma} = -\gamma\sigma\bar{\chi}_n$ (see (5.1)), while the corresponding eigenfunction w_n is given by (5.5),

$$w_n^{1,\sigma}(\chi) = D_A \operatorname{Ai}(\chi) + D_B \operatorname{Bi}(\chi), \quad \text{where } \chi \in [\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]. \tag{6.1}$$

Here, the coefficients D_A and D_B satisfy (5.6),

$$D_A \Gamma_{L,n}(\operatorname{Ai}) + D_B \Gamma_{L,n}(\operatorname{Bi}) = D_A \Gamma_{R,n}(\operatorname{Ai}) + D_B \Gamma_{R,n}(\operatorname{Bi}) = 0,$$

where $\Gamma_{L,n}(\cdot) = \Gamma(\cdot, \bar{\chi}_n)$ and $\Gamma_{R,n}(\cdot) = \Gamma\left(\cdot, \gamma^{-1} + \bar{\chi}_n\right)$. We treat the cases $\beta > 1$ and $0 < \beta < 1$ separately.

6.1. The case $\beta > 1$.

In this section, we select D_A and D_B so that (6.1) becomes

$$w_n^{1,\sigma}(\chi) = D_n \operatorname{Bi}(\chi) - \operatorname{Ai}(\chi), \quad \text{with } D_n = \frac{\Gamma_{L,n}(\operatorname{Ai})}{\Gamma_{L,n}(\operatorname{Bi})} = \frac{\Gamma_{R,n}(\operatorname{Ai})}{\Gamma_{R,n}(\operatorname{Bi})}. \tag{6.2}$$

Using this formula, we prove Lemma 6.1 and verify that ω_0 is of boundary layer type near $x = 1$.

6.1.1. The eigenfunction $w_0^{1,\sigma}$.

First, we show that $w_0^{1,\sigma}$ has no zeroes in the corresponding interval. Using Lemma A.1 and the estimates of Lemmas 5.4 for $\bar{\chi}_0$ and A.2 for $B_{0,\sigma}$, we estimate

$$D_0 = \left(\frac{\Delta_1^2}{2} + \bar{C}_0(\gamma)\right) \exp\left(-4\left(\frac{(\beta^2-1)^{3/4}}{3\gamma^{3/2}} + \sqrt{1 - \frac{1}{\beta^2}}\right)\right).$$

Here, $\Delta_1 = \beta + \sqrt{\beta^2 - 1}$ and $\left|\bar{C}_0(\gamma)\right| < c_0\sqrt{\gamma}$, for some $c_0 > 0$. Thus also, $D_0 > 0$.

It suffices to show that $w_0^{1,\sigma}$ is positive in this interval, and thus that $(w_0^{1,\sigma})' > 0$ everywhere on the interval and $w_0^{1,\sigma}(\bar{\chi}_0) > 0$. For $n = 0$, (6.2) yields $(w_0^{1,\sigma})'(\chi) =$

$D_0\,\mathrm{Bi}'(\chi) - \mathrm{Ai}'(\chi)$, while Lemma 5.4 shows that $[\bar{\chi}_0, \gamma^{-1} + \bar{\chi}_0] \subset \mathbf{R}_+$. Hence, $\mathrm{Bi}'(\chi) > 0$ and $\mathrm{Ai}'(\chi) < 0$ for all χ in this interval. Since $D_0 > 0$, we conclude that $(w_0^{1,\sigma})' > 0$, as desired. Next, we determine the sign of $w_0^{1,\sigma}(\bar{\chi}_0)$. This function is given in (6.2) with $n = 0$, while the definition of $\Gamma_{L,0}$ yields

$$\mathrm{Ai}(\bar{\chi}_0) = \Gamma_{L,0}(\mathrm{Ai}) + \beta^{-1}\sqrt{\gamma}\,\mathrm{Ai}'(\bar{\chi}_0) \quad \text{and} \quad \mathrm{Bi}(\bar{\chi}_0) = \Gamma_{L,0}(\mathrm{Bi}) + \beta^{-1}\sqrt{\gamma}\,\mathrm{Bi}'(\bar{\chi}_0).$$

Substituting into (6.2), we calculate $w_0^{1,\sigma}(\bar{\chi}_0) = \beta^{-1}\sqrt{\gamma}\,[D_0\,\mathrm{Bi}'(\bar{\chi}_0) - \mathrm{Ai}'(\bar{\chi}_0)]$. Thus, $w_0^{1,\sigma}(\bar{\chi}_0)$ is positive by our remarks on the signs of Bi', Ai', and D_0, and the proof is complete.

Next, we study the profile of the associated solution ω_0 to the original problem (1.14). Equations (4.1) and (5.1) yield

$$\omega_0(x) = \exp\left(\frac{\beta}{\gamma^{3/2}}\,x\right)\left[D_0\,\mathrm{Bi}(\gamma^{-1}x + \bar{\chi}_0) - \mathrm{Ai}(\gamma^{-1}x + \bar{\chi}_0)\right], \quad x \in [0,1].$$

Using the estimation of Lemma 5.4 for $\bar{\chi}_0$ and the estimations of Lemma A.1 for Ai and Bi, we find

$$\omega_0(x) = C_I(x)\left(x + \beta^2 - 1\right)^{-1/4}\exp\left(\frac{\beta}{\gamma^{3/2}}x\right)\sinh(\theta_1(x)), \quad x \in [0,1],$$

where $C_I(x) = C_{I,0} + C_{I,1}(x)$, $\sup_{[0,1]}|C_{I,1}(x)| < c_I\sqrt{\gamma}$, for some $c_I > 0$, and

$$\theta_1(x) = \frac{2}{3\gamma^{3/2}}\left[\left(x+\beta^2-1\right)^{3/2} - \left(\beta^2-1\right)^{3/2}\right] + \frac{2}{\beta}\left[\left(x+\beta^2-1\right)^{1/2} - \left(\beta^2-1\right)^{1/2}\right] + \log\Delta_1.$$

The first two terms on the right-hand side of the expression for ω_0 are bounded, while the other two correspond to localized concentrations (boundary layers) at $x = 1$. Thus, ω_0 also corresponds to a boundary layer of width $\mathcal{O}(\gamma^{3/2}) = \mathcal{O}(\sqrt{\varepsilon})$ at the same point.

6.1.2. The eigenfunctions $w_1^{1,\sigma}, \ldots, w_M^{1,\sigma}$. Next, we show that the eigenfunction $w_n^{1,\sigma}$ has n zeroes in $[\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]$, where $n = 1, \ldots, M$. The eigenfunction $w_n^{1,\sigma}$ is given by (6.2). Here also, Lemmas A.1 and 5.4 yield

$$D_n = \left(\frac{\Delta_2^2}{2} + \bar{C}_n(\gamma)\right)\exp\left(-\frac{4}{3\gamma^{3/2}} + 2\frac{|A_n|}{\sqrt{\gamma}} - \frac{2}{\beta}\right), \tag{6.3}$$

where $\Delta_2 = (\beta+1)^{1/2}(\beta-1)^{-1/2}$ and $\left|\bar{C}_n(\gamma)\right| < c_n\sqrt{\gamma}$, for some $c_n > 0$. Hence, $D_n > 0$.

First, we show that the function $w_n^{1,\sigma}$ has exactly $n-1$ zeroes in $[\bar{\chi}_n, 0]$. The estimate (6.3) and the fact that Bi is uniformly bounded on $[\bar{\chi}_n, 0]$ imply that, for all $0 < \gamma < \gamma_0$ (with γ_0 small enough), the functions $w_n^{1,\sigma}$ and $-\mathrm{Ai}$ are exponentially close in the W^1_∞-norm over that interval,

$$\left|\left|w_n^{1,\sigma} + \mathrm{Ai}\right|\right|_{[\bar{\chi}_n,0]} < c_n\exp\left(-\frac{4}{3\gamma^{3/2}} + 2\frac{|A_n|}{\sqrt{\gamma}}\right) \quad \text{for some } c_n > 0. \tag{6.4}$$

As a result, we may use an argument exactly analogous to the one used in the proof of Lemma 5.4 to show that $w_n^{1,\sigma}$ has at least $n-1$ distinct zeroes in $[\bar{\chi}_n, 0]$, each of which is exponentially close to one of $A_1, \ldots, A_{n-1}$. Observing that $\bar{\chi}_n$ is *algebraically*

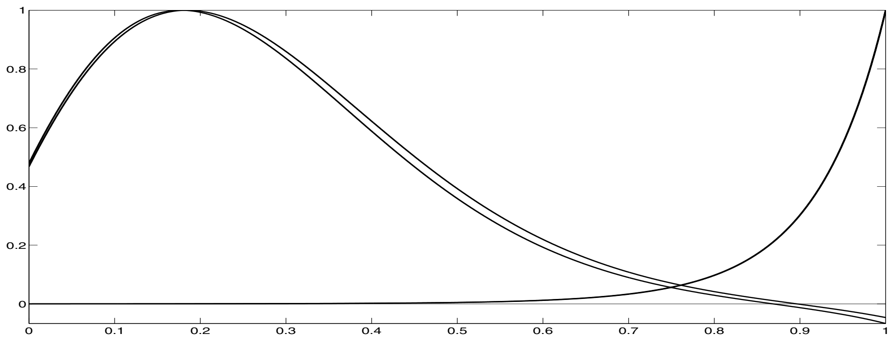


FIG. 6.1. *The eigenfunctions $w_0^{1,\sigma_L}, w_0^{1,\sigma_U}$ (always positive and coinciding within plotting accuracy) and $w_1^{1,\sigma_L}, w_1^{1,\sigma_U}$ (changing sign). Here, $a = 0.775$, $n_H = 0.667$, $\varepsilon = 0.001$, $\kappa = 1$, $\ell = 0.25$, and $j_H = 0.5$, which yields $\sigma_L = 0.1333$, $\sigma_U = 0.1457$ (and thus $\sigma_L < \sigma_U < a^2/4$), $0.0104 \leq \lambda_0 \leq 0.0222$, and $-0.0541 \leq \lambda_1 \leq -0.0512$. Note that $\lambda_1 < \lambda_0$ and that none of w_0^{1,σ_L} and w_0^{1,σ_U} has zeroes in $[0,1]$, while w_1^{1,σ_L} and w_1^{1,σ_U} have exactly one zero in the same interval.*

larger than A_n, by Lemmas 5.4 and A.2, while $w_n^{1,\sigma}$ is *exponentially* close to $-\operatorname{Ai}$, by estimate (6.4), we conclude that the zero of $w_n^{1,\sigma}$ close to A_n lies to the left of $\bar{\chi}_n$ (and hence outside $[\bar{\chi}_n, 0]$) and thus there are no other zeroes in $[\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]$.

It remains to show only that there is a unique zero of $w_n^{1,\sigma}$ in $[0, \gamma^{-1} + \bar{\chi}_n]$. We work as in section 6.1.1 and show that $w_n^{1,\sigma}$ is increasing and changes sign in that interval. First, we calculate $(w_n^{1,\sigma})'(\chi) = D_n \operatorname{Bi}'(\chi) - \operatorname{Ai}'(\chi) > 0$, where we have used that $\operatorname{Bi}'(\chi) > 0$, $\operatorname{Ai}'(\chi) < 0$, and $D_n > 0$. Also, $w_n^{1,\sigma}(0) < 0$ (by $\operatorname{Ai}(0) > 0$ and (6.4)) and, working as in section 6.1.1,

$$w_n^{1,\sigma}(\gamma^{-1} + \bar{\chi}_n) = \beta^{-1}\sqrt{\gamma}\,\left[D_n \operatorname{Bi}'(\gamma^{-1} + \bar{\chi}_n) - \operatorname{Ai}'(\gamma^{-1} + \bar{\chi}_n)\right] > 0.$$

This completes the proof.

6.2. The case $0 < \beta < 1$. In this section, we select D_A and D_B so that (6.1) becomes

$$w_n^{1,\sigma}(\chi) = \operatorname{Ai}(\chi) + D_n \operatorname{Bi}(\chi), \quad \text{with } D_n = -\frac{\Gamma_{L,n}(\operatorname{Ai})}{\Gamma_{L,n}(\operatorname{Bi})} = -\frac{\Gamma_{R,n}(\operatorname{Ai})}{\Gamma_{R,n}(\operatorname{Bi})}. \tag{6.5}$$

Using this formula, we prove Lemma 6.1 and verify that the profile of ω_0 has a spike around $x_\beta = \beta^2$.

We shall show that the eigenfunction $w_n^{1,\sigma}$ $(n = 0, \ldots, M)$ has n zeroes in $[\bar{\chi}_n, \gamma^{-1} + \bar{\chi}_n]$; see Figure 6.1. The proof is entirely analogous to that in section 6.1.2. Here also, the nth eigenvalue is $\mu_n^{1,\sigma} = -\gamma\sigma\bar{\chi}_n$, while the corresponding eigenfunction $w_n^{1,\sigma}$ is given by (6.5). The constant D_n may be estimated by

$$D_n = \left(\frac{\Delta_3^2}{2} + \hat{C}_n(\gamma)\right) \exp\left(-\frac{4}{3\gamma^{3/2}} + 2\frac{|A_{n+1}|}{\sqrt{\gamma}} - \frac{2}{\beta}\right), \tag{6.6}$$

where $\Delta_3 = \sqrt{1+\beta}/\sqrt{1-\beta}$ and $\left|\hat{C}_n\right| < c_n' \sqrt{\gamma}$ for some $c_n' > 0$. This is an estimate of the same type as (6.3) but with A_{n+1} replacing A_n. Thus, the estimate (6.4) holds here as well with the same change. Recalling that $\bar{\chi}_n$ is algebraically larger than

A_{n+1} (see Lemmas 5.4 and A.2), we conclude that $w_n^{1,\sigma}$ has n distinct zeroes, each of which is exponentially close to one of $A_1, \ldots, A_n$. Next, we show that $w_n^{1,\sigma} > 0$ in $[0, \gamma^{-1} + \bar{\chi}_n]$ and thus has no extra zeroes. First, $w_n^{1,\sigma}(\chi) = \operatorname{Ai}(\chi) + D_n \operatorname{Bi}(\chi)$. Now, $\operatorname{Bi}(\chi) > 0$ and $\operatorname{Ai}(\chi) > 0$, for all $\chi \in [0, \gamma^{-1} + \bar{\chi}_n]$, while $D_n > 0$ by (6.6). Hence, $w_n^{1,\sigma} > 0$, and the proof is complete.

Next, we examine the solution ω_0 associated with w_0. Working as in section 6.1.1, we calculate

$$\omega_0(x) = C_{II}(x) x^{-1/4} \exp\left(\frac{\beta}{\gamma^{3/2}} x\right) \cosh(\theta_2(x)), \quad x \in [0,1],$$

where $C_{II}(x) = C_{II,0} + C_{II,1}(x)$, $\sup_{[0,1]} |C_{II,1}(x)| < c_{II}\sqrt{\gamma}$ for some $c_{II} > 0$, and

$$\theta_2(x) = \frac{2}{3\gamma^{3/2}}\left(1 - x^{3/2}\right) - \left(\frac{|A_1|}{\sqrt{\gamma}} - \frac{1}{\beta}\right)(1 - \sqrt{x}) - \log \Delta_3.$$

The first two terms on the right-hand side of the expression for ω_0 are bounded, while the other two correspond to boundary layers at $x = 1$ and $x = 0$, respectively. A straightforward calculation shows that ω_0 corresponds to a spike of width $\mathcal{O}(\gamma^{3/4}) = \mathcal{O}(\varepsilon^{1/4})$ around the point x_β, where

$$\left|x_\beta - \left(\beta^2 + |A_1|\,\gamma\right)\right| < c\gamma^2 \quad \text{for some} \quad c > 0. \tag{6.7}$$

We remark that x_β does *not* correspond to the position of the DCM for the problem (1.14) involving the function f. This information is obtained in the next section, instead, through a WKB analysis.

7. The WKB approximation.

In the previous sections, we derived strict bounds for the eigenvalues $\mu_1, \ldots, \mu_M$ of $\mathcal{L}$ and summarized them in Theorem 2.1. In this section, we use the WKB method to derive explicit (albeit asymptotic) formulas for these eigenvalues. The outcome of this analysis has already been summarized in section 2.1.

7.1. The case $A < \sigma_L$.

7.1.1. WKB formulas for w.

The eigenvalue problem (4.2) reads

$$\varepsilon w_{xx} = (F(x) - \mu)\, w, \quad \text{with} \quad \mathcal{G}(w, 0) = \mathcal{G}(w, 1) = 0. \tag{7.1}$$

Since we are interested in the regime $\sigma_L > A$, Lemma 4.3 states that the eigenvalues $\mu_0, \ldots, \mu_M$ lie in a $\mathcal{O}(\varepsilon^{1/3})$ region to the right of zero. Thus, for any $0 \le n \le M$,

$$F(x) < \mu_n \quad \text{for } x \in [0, \bar{x}_n), \quad \text{and} \quad F(x) > \mu_n \quad \text{for } x \in (\bar{x}_n, 1].$$

Here, $\bar{x}_n$ corresponds to a *turning point*, i.e., $F(\bar{x}_n) = \mu_n$, and it is given by the formula

$$\bar{x}_n = \frac{1}{\kappa} \log \frac{1 + \mu_n(1 + \eta_H)(1 + j_H^{-1})}{1 - \mu_n(1 + \eta_H)(1 + j_H)}. \tag{7.2}$$

Lemmas 4.3 and A.2 suggest that the eigenvalue μ_n may be expanded asymptotically in powers of $\varepsilon^{1/6}$ starting with $\mathcal{O}(\varepsilon^{1/3})$ terms, $\mu_n = \sum_{\ell=2}^{\infty} \varepsilon^{\ell/6} \mu_{n,\ell}$. Thus, we also find

$$\bar{x}_n = \varepsilon^{1/3} \sigma_0^{-1} \mu_{n,2} + \varepsilon^{1/2} \sigma_0^{-1} \mu_{n,3} + \mathcal{O}\left(\varepsilon^{2/3}\right), \quad \text{where } \sigma_0 = F'(0). \tag{7.3}$$

The solution in the region $(\bar{x}_n, 1]$, where $F(x) - \mu_n > 0$, can be determined using standard formulas (see [2, section 10.1]),

$$w_n(x) = [F(x)-\mu_n]^{-1/4}\left[C_a \exp^{-\int_{\bar{x}_n}^{x} \sqrt{(F(s)-\mu_n)/\varepsilon}\, ds} + C_b\, \mathrm{e}^{\int_{\bar{x}_n}^{x} \sqrt{(F(s)-\mu_n)/\varepsilon}\, ds}\right]. \tag{7.4}$$

Here, C_a and C_b are arbitrary constants, to leading order in ε. (Higher order terms in the asymptotic expansions of C_a and C_b generally depend on x; see [2] for details.) Using this information and the asymptotic expansion for μ_n, we may determine the principal part of the solution w_n,

$$w_{n,0}(x) = [F(x)]^{-1/4}\left[C_{a,0}\, \mathrm{e}^{-\theta_3(x)} + C_{b,0}\, \mathrm{e}^{\theta_3(x)}\right], \tag{7.5}$$

for arbitrary constants $C_{a,0}$ and $C_{b,0}$ and where

$$\theta_3(x) = \frac{1}{\varepsilon^{1/2}}\int_0^x \sqrt{F(s)}\, ds - \frac{1}{\varepsilon^{1/6}}\frac{\mu_{n,2}}{2}\int_0^x \frac{ds}{\sqrt{F(s)}} + \frac{\mu_{n,2}}{\sqrt{\sigma_0}} - \frac{2}{3}\sqrt{\sigma_0} - \frac{\mu_{n,3}}{2}\int_0^x \frac{ds}{\sqrt{F(s)}}. \tag{7.6}$$

To determine the solution in $[0, \bar{x}_n)$, we change the independent variable through

$$x = \varepsilon^{1/3}\sigma_0^{-1/3}(\chi - \bar{\chi}_n), \quad \text{where} \quad \bar{\chi}_n = -\sigma_0^{1/3}\,\varepsilon^{-1/3}\,\bar{x}_n = -\sigma_0^{-2/3}\mu_{n,2} + \mathcal{O}\left(\sqrt{\varepsilon}\right) < 0, \tag{7.7}$$

and expand $F(x) - \mu_n = F(x) - F(\bar{x}_n)$ asymptotically:

$$F(x) - F(\bar{x}_n) = F(\varepsilon^{1/3}\sigma_0^{-1/3}(\chi - \bar{\chi}_n)) - F(-\varepsilon^{1/3}\sigma_0^{-1/3}\bar{\chi}_n) = \varepsilon^{1/3}\sigma_0^{2/3}\chi + \mathcal{O}(\sqrt{\varepsilon}). \tag{7.8}$$

As a result, (7.1) becomes the Airy equation $(w_n)_{\chi\chi} = \chi w_n$, to leading order, whence

$$w_{n,0}(\chi) = D_{a,0}\,\mathrm{Ai}(\chi) + D_{b,0}\,\mathrm{Bi}(\chi), \quad \text{with } \chi \in (-\sigma_0^{-2/3}\mu_{n,2}, 0]. \tag{7.9}$$

7.1.2. Boundary conditions for the WKB solution. Next, we determine the coefficients appearing in (7.5) and (7.9). Formula (7.5) represents the solution in the region $(\bar{x}_n, 1]$, and thus it must satisfy the boundary condition $\mathcal{G}\,(w_n, 1) = 0$. Using (4.3), we find, to leading order,

$$C_{a,0}\,(a + 2\sqrt{\sigma_1})\,\mathrm{e}^{-\theta_3(x)} + C_{b,0}\,(a - 2\sqrt{\sigma_1})\,\mathrm{e}^{\theta_3(x)} = 0, \quad \text{where } \sigma_1 = F(1). \tag{7.10}$$

Next, the formula given in (7.9) is valid for $\chi \in (-\sigma_0^{-2/3}\mu_{n,2}, 0]$ (equivalently, for $x \in [0, \bar{x}_n)$), and thus it must satisfy the boundary condition $\mathcal{G}\,(w, 0) = 0$. Recasting the formula for $\mathcal{G}$ given in (4.3) in terms of χ, we obtain to leading order the equation

$$D_{a,0}\,\mathrm{Ai}\left(-\sigma_0^{-2/3}\mu_{n,2}\right) + D_{b,0}\,\mathrm{Bi}\left(-\sigma_0^{-2/3}\mu_{n,2}\right) = 0. \tag{7.11}$$

Finally, (7.5) and (7.9) must also match in an intermediate length scale to the right of $x = \bar{x}_n$ (equivalently, of $\chi = 0$). To this end, we set $\psi = \varepsilon^d\,(x - \bar{x}_n)$, where $1/5 < d < 1/3$ [2, section 10.4], and recast (7.5) in terms of ψ. We find, to leading order and for all $\mathcal{O}(1)$ and positive values of ψ,

$$w_{n,0}(x(\psi)) = \varepsilon^{-d/4}\,\sigma_0^{-1/4}\,\psi^{-1/4}\left[C_{a,0}\,\mathrm{e}^{-\theta_4(\psi)-\sigma_0^{-1}(\mu_{n,2})^{3/2}} + C_{b,0}\,\mathrm{e}^{\theta_4(\psi)+\sigma_0^{-1}(\mu_{n,2})^{3/2}}\right],$$

where $\theta_4(\psi) = (2/3)\,\varepsilon^{(3d-1)/2}\,\sqrt{\sigma_0}\,\psi^{3/2}$. Similarly, (7.9) yields

$$w_{n,0}(\chi(\psi)) = \varepsilon^{1/12-d/4}\,\sigma_0^{-1/12}\,\pi^{-1/2}\,\psi^{-1/4}\left[\frac{D_{a,0}}{2}\,\mathrm{e}^{-\theta_4(\psi)} + D_{b,0}\,\mathrm{e}^{\theta_4(\psi)}\right].$$

The matching condition around the turning point then gives

$$(7.12)\quad C_{a,0} = \varepsilon^{1/12}\frac{\sigma_0^{1/6}}{2\sqrt{\pi}}\mathrm{e}^{\sigma_0^{-1}(\mu_{n,2})^{3/2}}D_{a,0} \quad \text{and} \quad C_{b,0} = \varepsilon^{1/12}\frac{\sigma_0^{1/6}}{\sqrt{\pi}}\mathrm{e}^{-\sigma_0^{-1}(\mu_{n,2})^{3/2}}D_{b,0}.$$

7.1.3. The eigenvalues $\boldsymbol{\mu_0, \dots, \mu_n}$. The linear system (7.10)–(7.12) has a nontrivial solution if and only if the determinant corresponding to it vanishes identically:

$$\begin{aligned} &2\,(a - 2\sqrt{\sigma_1})\,\mathrm{e}^{\theta_3(1) - \sigma_0^{-1}(\mu_{n,2})^{3/2}}\,\mathrm{Ai}(\sigma^{-2/3}\mu_{n,2}) \\ &\qquad + (a + 2\sqrt{\sigma_1})\,\mathrm{e}^{-\theta_3(1) + \sigma_0^{-1}(\mu_{n,2})^{3/2}}\,\mathrm{Bi}(\sigma^{-2/3}\mu_{n,2}) = 0. \end{aligned}$$

Since $\sigma_1 \geq \sigma_L$ by Lemma 2.1 and $\sigma_L > A$ by assumption, $a - 2\sqrt{\sigma_1}$ is $\mathcal{O}(1)$ and negative. Also, $\theta_3(1)$ is $\mathcal{O}(1)$ and positive by (7.6). Thus, the determinant condition reduces to $\mathrm{Ai}(\sigma^{-2/3}\mu_{n,2}) = 0$, whence $\mu_{n,2} = -\sigma_0^{2/3}\,A_{n+1} = \sigma_0^{2/3}\,|A_{n+1}| > 0$. Hence, we find for the eigenvalues of (1.14)

$$(7.13)\qquad \lambda_n = \lambda^* - \varepsilon^{1/3}\sigma_0^{2/3}\,|A_{n+1}| + \mathcal{O}(\sqrt{\varepsilon}).$$

Working in a similar way, we find $\mu_{n,3} = -2\sigma_0/a$.

Recalling that $\sigma_0 = F'(0) = -f'(0)$ by (2.1) and Lemma 2.1 (see also Figure 2.3), we find that the WKB formula (7.13) coincides—up to and including terms of $\mathcal{O}(1)$ and $\mathcal{O}(\varepsilon^{1/3})$—(a) for $0 < j_H < j_H^{(2)}$, with the rigorous lower bound for λ_n derived in Theorem 2.1, and (b) for $j_H > 1$, with the rigorous upper bound for λ_n derived in the same theorem. For the remaining values of j_H, (7.13) yields a value for λ_n which lies in between the upper and lower bounds derived in Theorem 2.1—indeed, in that case, $\sigma_L < F'(0) < \sigma_U$; see Figure 2.3.

7.1.4. The eigenfunctions $\boldsymbol{w_0, \dots, w_n}$. Finally, one may determine the constants C_a, C_b, D_a, and D_b corresponding to the eigenfunction w_n, and thus also w_n itself, through (7.10)–(7.12). The principal part of w_n is given by the formula

$$(7.14)\qquad w_{n,0}(x) = \begin{cases} \mathrm{Ai}\left(A_{n+1} + \varepsilon^{-1/3}\sigma_0^{1/3}x\right) & \text{for } x \in [0, \varepsilon^{1/3}\sigma_0^{-1/3}\,|A_{n+1}|), \\ C\,[F(x)]^{-1/4}\cosh\Theta(x) & \text{for } x \in (\varepsilon^{1/3}\sigma_0^{-1/3}\,|A_{n+1}|\,, 1]. \end{cases}$$

Here,

$$(7.15)\quad C = \varepsilon^{1/12}\,\frac{\sigma_0^{1/6}}{2\sqrt{\pi}}\,\Delta_4\,\mathrm{e}^{|A_{n+1}|^{3/2} - \Theta_3(1)}, \quad \text{where } \Delta_4 = \left(\frac{\sqrt{\sigma_1} + \sqrt{A}}{\sqrt{\sigma_1} - \sqrt{A}}\right)^{1/2},$$

$$(7.16)\qquad \Theta(x) = \varepsilon^{-1/2}\int_x^1 \sqrt{F(s)}\,ds - \left(\varepsilon^{-1/6}\,\frac{\sigma_0^{2/3}\,|A_{n+1}|}{2} - \frac{\sigma_0}{a}\right)\int_x^1 \frac{ds}{\sqrt{F(s)}} + \log\Delta_4.$$

Recalling (4.1), we find

$$(7.17)\qquad \omega_{n,0}(x) = \begin{cases} \mathrm{e}^{\sqrt{A/\varepsilon}\,x}\,\mathrm{Ai}\left(A_{n+1} + \varepsilon^{-1/3}\sigma_0^{1/3}x\right) & \text{for } x \in [0, \varepsilon^{1/3}\sigma_0^{-1/3}\,|A_{n+1}|), \\ C\,[F(x)]^{-1/4}\,\mathrm{e}^{\sqrt{A/\varepsilon}\,x}\cosh\Theta(x) & \text{for } x \in (\varepsilon^{1/3}\sigma_0^{-1/3}\,|A_{n+1}|\,, 1]. \end{cases}$$

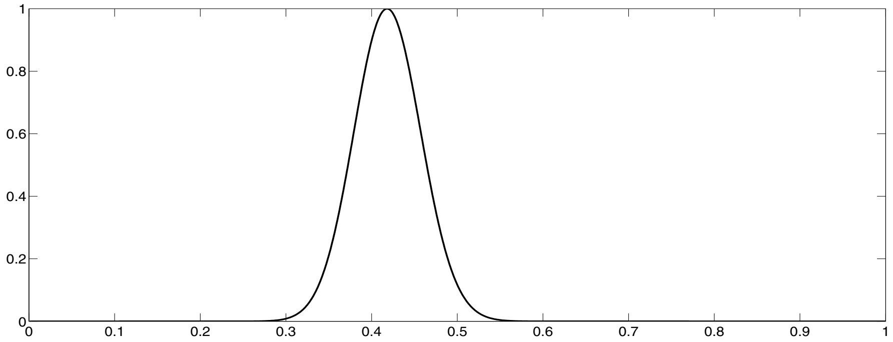


FIG. 7.1. *The eigenfunction ω_0 as given by* (7.17). *Here,* $a = 0.5$, $n_H = 0.667$, $\varepsilon = 2 \cdot 10^{-7}$, $\kappa = 1$, *and* $j_H = 0.5$. *The eigenfunction has been scaled so that its maximum value is equal to one.*

A straightforward calculation shows that ω_0 corresponds to a spike around the point

$$x_{\mathrm{DCM}} = x_{\mathrm{DCM},0} + \mathcal{O}(\varepsilon^{1/3}), \tag{7.18}$$

where $x_{\mathrm{DCM},0}$ is the unique solution to $F(x_{\mathrm{DCM},0}) = A = a^2/4$; see also Figure 7.1, where ω_0 is plotted for specific parameter values. Thus, $\omega_{0,0}$ indeed corresponds to a DCM. Furthermore, the location of the maximum phytoplankton concentration is expressed explicitly by this equation in terms of the rescaled biological parameters κ, η_H, j_H, and a.

7.2. The case $A > \sigma_U$. To obtain the eigenvalues and their corresponding eigenfunctions in this case, we work as in the preceding section. Here also, the eigenvalue problem (4.2) has the form (7.1). Since $A > \sigma_U$, the eigenvalue μ_0 is $\mathcal{O}(1)$ and negative, while $\mu_1, \ldots, \mu_M$ are $\mathcal{O}(\varepsilon^{1/3})$ and positive; see Lemma 4.3. Due to the qualitative difference between μ_0 and the eigenvalues of higher order, we consider them separately.

We start with the case $1 \leq n \leq M$. Then, for each such n, the eigenvalue problem (7.1) has a unique turning point $\bar{x}_n$ given by (7.2), and the analysis presented in the preceding section applies here also. The formulas for μ_n and ω_n, $1 \leq n \leq M$, are identical to those of the preceding section, with the sole modification that A_n in (7.13)–(7.16) must be replaced by A_{n-1}. This completes the analysis for the case $1 \leq n \leq M$.

Next, we treat the case $n = 0$. Since $\mu_0 < 0 < F(x)$ for all $x \in [0, 1]$, the eigenvalue problem (7.1) corresponding to μ_0 has no turning points. Thus, the WKB formula (7.4), with $n = 0$ and $\bar{x}_n$ replaced by zero, is valid for all $x \in [0, 1]$. Lemmas 4.3 and A.2 suggest that μ_0 may be expanded asymptotically as $\mu_0 = \sum_{\ell=0}^{\infty} \varepsilon^{\ell/2} \mu_{0,\ell}$. Using this expansion, we calculate the principal part of w_0,

$$w_{0,0}(x) = [F(x) - \mu_{0,0}]^{-1/4} \left[C_{a,0}\, \mathrm{e}^{-\theta_5(x)} + C_{b,0}\, \mathrm{e}^{\theta_5(x)} \right], \tag{7.19}$$

where $C_{a,0}$ and $C_{b,0}$ are arbitrary constants and

$$\theta_5(x) = \frac{1}{\varepsilon^{1/2}} \int_0^x \sqrt{F(s) - \mu_{0,0}}\, ds - \frac{\mu_{0,1}}{2} \int_0^x \frac{ds}{\sqrt{F(s) - \mu_{0,0}}}. \tag{7.20}$$

Next, recalling the boundary conditions $\mathcal{G}\,(w,0) = \mathcal{G}\,(w,1) = 0$, we obtain, to leading order,

$$\begin{aligned} C_{a,0}\,(a + 2\sqrt{-\mu_{0,0}}) + C_{b,0}\,(a - 2\sqrt{-\mu_{0,0}}) &= 0, \\ C_{a,0}\,(a + 2\sqrt{\sigma_1 - \mu_{0,0}})\,\mathrm{e}^{-\theta_5(1)} + C_{b,0}\,(a - 2\sqrt{\sigma_1 - \mu_{0,0}})\,\mathrm{e}^{\theta_5(1)} &= 0, \end{aligned} \tag{7.21}$$

where we recall that $\sigma_1 = F(1)$. Here, $\theta_5(1)$ is $\mathcal{O}(1)$ and positive by (7.20), while $a + 2\sqrt{-\mu_{0,0}} > 0$. Thus, we obtain $\mu_{0,0} = F(1) - A$, to leading order, whence

$$\lambda_{0,0} = f(1) - \ell.$$

This is precisely (2.8). Using this formula, one may also determine $C_{a,0}$ and $C_{b,0}$ to obtain $w_{0,0}$,

$$w_{0,0}(x) = \,[F(x) - \mu_{0,0}]^{-1/4} \sinh \Phi(x), \tag{7.22}$$

for $x \in [0,1]$ and up to a multiplicative constant. Here,

$$\Phi(x) = \frac{1}{\varepsilon^{1/2}} \int_0^x \sqrt{F(s) - \mu_{0,0}}\, ds - \frac{\mu_{0,1}}{2} \int_0^x \frac{ds}{\sqrt{F(s) - \mu_{0,0}}} + \log \Delta_5,$$

where

$$\Delta_5 = \beta_1 + \sqrt{\beta_1^2 - 1} \quad \text{and} \quad \beta_1 = \frac{\sqrt{A}}{F(1)}.$$

Recalling (4.1), we find

$$\omega_{0,0}(x) = [F(x) - \mu_{0,0}]^{-1/4}\, \mathrm{e}^{ax/2\sqrt{\varepsilon}} \sinh \Phi(x) \quad \text{for } x \in [0,1].$$

The profile of ω_0 corresponds to a boundary layer at the point $x = 1$.

7.3. The transitional regime $\boldsymbol{\sigma_L < A < \sigma_U}$.

Equations (2.9) and (2.8) may be used to derive information for the transitional regime $\sigma_L < A < \sigma_U$ (see Theorem 2.1 and the discussion in section 2). In particular, the transition between the case where λ_0 is associated with a boundary layer (in biological terms, with a BL) and the case where it is associated with a spike (that is, with a DCM) occurs, to leading order, when $f(1) - \ell = \lambda^*$. Recalling (2.5), we rewrite this equation as

$$F(1) = f(0) - f(1) = A. \tag{7.23}$$

This condition reduces, to leading order, to $A = \sigma_U$ for $0 < j_H \leq j_H^{(1)}$, and to $A = \sigma_L$ for $j_H \geq j_H^{(2)}$. For $j_H^{(1)} < j_H < j_H^{(2)}$, this transitional value of A lies between σ_U and σ_L; see section 2 and Figure 2.3.

Appendix A. Basic properties of the Airy functions.

In this section, we summarize some properties of the Airy functions Ai and Bi which we use repeatedly.

LEMMA A.1. *Let $p > 0$ and q be real numbers. Then,*

$$\begin{aligned} &\Gamma\left(\mathrm{Ai}, \gamma^{-1} p + q\right) \\ &\quad = (\pi^{-1/2}/2)\left(\gamma\, p^{-1}\right)^{1/4} \exp\left(-\,(2/3)\left(\gamma^{-1} p\right)^{3/2} - q\left(\gamma^{-1} p\right)^{1/2}\right) \\ &\qquad \cdot \left[\left(1 + \beta^{-1}\sqrt{p}\right)\left(1 - (q^2/4)\left(\gamma\, p^{-1}\right)^{1/2} + (q/4)\left(q^3/8 - 1\right)\gamma\, p^{-1}\right)\right. \\ &\qquad \left. -\,(1/48)\left(5 - 5q^3 + q^6/8 - \left(43 - q^3 - q^6/8\right)\beta^{-1}\sqrt{p}\right)\left(\gamma\, p^{-1}\right)^{3/2}\right], \quad \gamma \downarrow 0, \end{aligned}$$

$$
\begin{aligned}
&\Gamma\left(\mathrm{Bi},\gamma^{-1}p+q\right)\\
&\quad= \pi^{-1/2}\left(\gamma\, p^{-1}\right)^{1/4}\exp\left((2/3)\left(\gamma^{-1}p\right)^{3/2}+q\left(\gamma^{-1}p\right)^{1/2}\right)\\
&\qquad\cdot\Big[\left(1-\beta^{-1}\sqrt{p}\right)\left(1+\left(q^2/4\right)\left(\gamma\, p^{-1}\right)^{1/2}+(q/4)\left(q^3/8-1\right)\gamma\, p^{-1}\right)\\
&\qquad+(1/48)\left(5-5q^3+q^6/8+\left(43-q^3-q^6/8\right)\beta^{-1}\sqrt{p}\right)\left(\gamma\, p^{-1}\right)^{3/2}\Big], \quad \gamma\downarrow 0,
\end{aligned}
$$

where the remainders of $\mathcal{O}(\gamma^2)$ *were omitted from within the square brackets.*

Proof. We derive only the first of these asymptotic expansions. The second one is derived in an entirely analogous manner. Definition (5.2) yields

$$
\Gamma\left(\mathrm{Ai},\gamma^{-1}\,p+q\right)=\mathrm{Ai}\left(\gamma^{-1}\,p+q\right)-\sqrt{\gamma}\,\beta^{-1}\,\mathrm{Ai}'\left(\gamma^{-1}\,p+q\right).
$$

Then, we recall the standard asymptotic expansions [2]

$$
\begin{aligned}
\mathrm{Ai}(z)&=\left(\pi^{-1/2}\,z^{-1/4}/2\right)\exp\left(-(2/3)z^{3/2}\right)\left[1-(5/48)\,z^{-3/2}+\mathcal{O}(z^{-3})\right], \quad z\uparrow\infty,\\
\mathrm{Ai}'(z)&=-\left(\pi^{-1/2}\,z^{1/4}/2\right)\exp\left(-(2/3)z^{3/2}\right)\left[1+(7/48)\,z^{-3/2}+\mathcal{O}(z^{-3})\right], \quad z\uparrow\infty,
\end{aligned}
$$

$$
\left(\gamma^{-1}p+q\right)^r=p^r\gamma^{-r}+\sum_{k=1}^{\infty}\frac{1}{k!}\left(\prod_{j=0}^{k-1}(r-j)\right)p^{r-k}q^k\,\gamma^{k-r}.
$$

The desired equation now follows by combining these asymptotic expansions. □

Corollary A.1. *Let* p *and* q *be as in Lemma* A.1. *Then, for* $\gamma\downarrow 0$,

$$
\begin{aligned}
\Gamma\left(\mathrm{Ai}',\gamma^{-1}\,p+q\right)&=-\left(\pi^{-1/2}/2\right)\left(\gamma^{-1}\,p\right)^{1/4}\exp\left(-(2/3)\left(\gamma^{-1}p\right)^{3/2}-q\left(\gamma^{-1}p\right)^{1/2}\right)\\
&\quad\cdot\Big[\left(1+\beta^{-1}\sqrt{p}\right)\left(1-\left(q^2/4\right)\left(\gamma\, p^{-1}\right)^{1/2}\right)\\
&\quad+(q/4)\left(\left(q^3/8-1\right)+\left(q^3/8+3\right)\beta^{-1}\sqrt{p}\right)\gamma\, p^{-1}\\
&\quad-(1/48)\left(-19+q^3+q^6/8+\left(-7+7q^3+q^6/8\right)\beta^{-1}\sqrt{p}\right)\left(\gamma\, p^{-1}\right)^{3/2}\Big],\\
\Gamma\left(\mathrm{Bi}',\gamma^{-1}\,p+q\right)&=\pi^{-1/2}\left(\gamma^{-1}\,p\right)^{1/4}\exp\left((2/3)\left(\gamma^{-1}\,p\right)^{3/2}+q\left(\gamma^{-1}\,p\right)^{1/2}\right)\\
&\quad\cdot\Big[\left(1-\beta^{-1}\sqrt{p}\right)\left(1+\left(q^2/4\right)\left(\gamma p^{-1}\right)^{1/2}\right)\\
&\quad+(q/4)\left(\left(q^3/8-1\right)-\left(q^3/8+3\right)\beta^{-1}\sqrt{p}\right)\gamma\, p^{-1}\\
&\quad+(1/48)\left(-19+q^3+q^6/8-\left(-7+7q^3+q^6/8\right)\beta^{-1}\sqrt{p}\right)\left(\gamma\, p^{-1}\right)^{3/2}\Big],
\end{aligned}
$$

where the remainders of $\mathcal{O}(\gamma^2)$ *were omitted from within the square brackets.*

Proof. Definition (5.2) and the identities $\mathrm{Ai}''(z)=z\,\mathrm{Ai}(z)$ and $\mathrm{Bi}''(z)=z\,\mathrm{Bi}(z)$ yield

$$
\begin{aligned}
\Gamma\left(\mathrm{Ai}',\gamma^{-1}p+q\right)&=\mathrm{Ai}'\left(\gamma^{-1}p+q\right)-\sqrt{\gamma}\,\beta^{-1}\left(\gamma^{-1}\,p+q\right)\mathrm{Ai}\left(\gamma^{-1}\,p+q\right),\\
\Gamma\left(\mathrm{Bi}',\gamma^{-1}+\bar{\chi}\right)&=\mathrm{Bi}'\left(\gamma^{-1}\,p+q\right)-\sqrt{\gamma}\,\beta^{-1}\left(\gamma^{-1}\,p+q\right)\mathrm{Bi}\left(\gamma^{-1}\,p+q\right).
\end{aligned}
$$

The desired result now follows from Lemma A.1. □

Lemma A.2. *The function* $\Gamma\left(\mathrm{Ai},\bar{\chi}\right)$ *has no positive roots. Further, for any* $M\in\mathbf{N}$, *there is an* $\varepsilon_0>0$ *such that, for all* $0<\varepsilon<\varepsilon_0$, $\Gamma\left(\mathrm{Ai},\bar{\chi}\right)$ *has roots* $A'_{M,\sigma}<\cdots<A'_{1,\sigma}<0$ *satisfying*

$$
\left|A'_{n,\sigma}-\left(A_n+\sqrt{\gamma}\,\beta^{-1}\right)\right|<c_a\,\gamma \quad \textit{for some} \quad c_a>0.
$$

Here, $A_n < 0$ is the nth root of Ai *(see Figure* 2.1*), and β, γ are given in* (2.3)*. For $\beta > 1$ (equivalently, for $0 < \sigma < a^2/4$), the function $\Gamma\left(\mathrm{Bi}, \gamma^{-1}(1+\bar{\psi})\right)$ defined in* (2.4) *has a root $B_{0,\sigma} > 0$ satisfying*

$$\left|B_{0,\sigma} - \left(\beta^2 - 1 + 2\,\gamma^{3/2}\,\beta^{-1}\right)\right| < c_b\,\gamma^3 \quad \textit{for some} \quad c_b > 0.$$

Proof. The fact that there exist no positive roots of $\Gamma\,(\mathrm{Ai}, \bar{\chi})$ is immediate by the definition of $\Gamma\,(\mathrm{Ai}, \bar{\chi})$ (see (2.4)) and the fact that $\mathrm{Ai}(\bar{\chi}) > 0$ and $\mathrm{Ai}'(\bar{\chi}) < 0$ for all $\bar{\chi} > 0$.

Next, the existence of M discrete and negative roots may be proved as follows. Fix $|A_M| < X < |A_{M+1}|$ and let $I_1, \ldots, I_M$ be disjoint intervals around $A_1, \ldots, A_M$, respectively. Definition (2.4) implies that $\Gamma\,(\mathrm{Ai}, \cdot)$ is $\mathcal{O}(\sqrt{\gamma})$ close to Ai over $[-X, 0]$ in the norm introduced in (5.10). Thus, for all $0 < \gamma < \gamma_0$ and γ_0 small enough, $\Gamma\,(\mathrm{Ai}, \bar{\chi})$ has M distinct roots $A'_{1,\sigma} \in I_1, \ldots, A'_{M,\sigma} \in I_M$ in $[-X, 0]$. That these are ordered as $A'_{M,\sigma} < \cdots < A'_{1,\sigma}$ follows from $A_{M,\sigma} < \cdots < A_{1,\sigma}$ and the fact that $I_1, \ldots, I_M$ were chosen to be disjoint. The bounds on $A'_{1,\sigma}, \ldots, A'_{M,\sigma}$ may be derived by writing $A'_{n,\sigma} = \sum_{\ell \geq 0} \varepsilon^{\ell/6}\, a^{(\ell)}_{n,\sigma}$, substituting into the equation $\Gamma\,(\mathrm{Ai}, \bar{\chi}) = 0$, and expanding asymptotically.

The existence of $B_{0,\sigma} > 0$ and the bound on it may be established using Lemma A.1 (with $p = 1 + \bar{\psi}$ and $q = 0$). □

Appendix B. Proof of Lemma 5.2. Using definition (5.8), we calculate

$$\mathcal{A}(\bar{\chi}) - \Gamma\,(\mathrm{Ai}, \bar{\chi}) = -\frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)}{\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})}\Gamma\,(\mathrm{Bi}, \bar{\chi})\,. \tag{B.1}$$

To estimate the fraction on the right-hand side, we apply standard theory for Airy functions [2]; see Appendix A. Using Lemma A.1 (with $p = 1$ and $q = \bar{\chi}$), we find that

$$\sup_{\bar{\chi} \in [X,0]} \left| \exp\left(\frac{4}{3\gamma^{3/2}} + \frac{2\bar{\chi}}{\gamma^{1/2}}\right) \frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)}{\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})} - \frac{1}{2}\frac{\beta + 1}{\beta - 1} \right| < c_1 \sqrt{\gamma},$$

for some $c_1 > 0$ and γ small enough. Therefore,

$$\sup_{\bar{\chi} \in [X,0]} \left| \frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)}{\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})} \right| < c_2 \exp\left(-\frac{4 + 6\,\gamma\,X}{3\gamma^{3/2}}\right), \tag{B.2}$$

for some $c_2 > 0$. Next, $\sup_{[X,0]} |\Gamma\,(\mathrm{Bi}, \cdot)| \leq c_3$ for some $c_3 > 0$, since Bi and Bi$'$ are uniformly bounded over $[X, 0]$. Combining these estimates, we find

$$\sup_{\bar{\chi} \in [X,0]} |\mathcal{A}(\bar{\chi}) - \Gamma\,(\mathrm{Ai}, \bar{\chi})| < c_4 \exp\left(-\frac{4 + 6\,\gamma\,X}{3\gamma^{3/2}}\right), \tag{B.3}$$

for some $c_4 > 0$ and for all γ small enough.

Next, differentiating (B.1), we calculate

$$\begin{aligned} \mathcal{A}'(\bar{\chi}) - \Gamma(\mathrm{Ai}', \bar{\chi}) = {} & \left(\frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right) \Gamma\left(\mathrm{Bi}', \gamma^{-1} + \bar{\chi}\right)}{[\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})]^2} - \frac{\Gamma\left(\mathrm{Ai}', \gamma^{-1} + \bar{\chi}\right)}{\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})} \right) \Gamma\,(\mathrm{Bi}, \bar{\chi}) \\ & - \frac{\Gamma\left(\mathrm{Ai}, \gamma^{-1} + \bar{\chi}\right)}{\Gamma\,(\mathrm{Bi}, \gamma^{-1} + \bar{\chi})} \Gamma\left(\mathrm{Bi}', \bar{\chi}\right). \end{aligned} \tag{B.4}$$

Using Lemma A.1, we may bound the term in parentheses by

$$\frac{c_1'}{\sqrt{\gamma}} \exp\left(-\frac{4+6\,\gamma\,X}{3\,\gamma^{3/2}}\right),$$

for some $c_1' > 0$. Next, $\Gamma\left(\mathrm{Bi}, \bar{\chi}\right)$ was uniformly bounded by a constant c_3 above. Also, the term $\Gamma\left(\mathrm{Bi}', \bar{\chi}\right)$ may be bounded by a constant c_3', since

$$\Gamma\left(\mathrm{Bi}', \bar{\chi}\right) = \mathrm{Bi}'(\bar{\chi}) - \sqrt{\gamma}\,\beta\,\mathrm{Bi}''(\bar{\chi}) = \mathrm{Bi}'(\bar{\chi}) - \sqrt{\gamma}\,\beta\,\bar{\chi}\,\mathrm{Bi}(\bar{\chi}),$$

and the term multiplying it in (B.4) was bound in (B.2). These inequalities yield, then,

$$||\mathcal{A}'(\cdot) - \mathrm{Ai}'(\cdot)||_{[X,0]} < c_2'\,\gamma^{-1/2} \exp\left(-\frac{4+6X\gamma}{3\gamma^{3/2}}\right), \tag{B.5}$$

for some $c_2' > 0$ and for all γ small enough. Equation (5.11) follows now from (B.3) and (B.5). □

Appendix C. Proof of Lemma 5.3. Definition (5.9) yields

$$\mathcal{B}(\gamma^{-1}\bar{\psi}) - \Gamma\left(\mathrm{Bi}, \gamma^{-1}(1+\bar{\psi})\right) = -\frac{\Gamma\left(\mathrm{Bi}, \gamma^{-1}\bar{\psi}\right)}{\Gamma\left(\mathrm{Ai}, \gamma^{-1}\bar{\psi}\right)}\Gamma\left(\mathrm{Ai}, \gamma^{-1}(1+\bar{\psi})\right). \tag{C.1}$$

To estimate the right-hand side, we work as in Appendix B. Using Lemma A.1 twice (once with $p = \bar{\psi}$, $q = 0$ and once with $p = 1+\bar{\psi}$, $q = 0$), we obtain

$$\begin{aligned} &\sup_{\bar{\psi}\in[\Psi_R,\Psi_L]} \left| E(\gamma^{-1}(1+\bar{\psi}))\,\frac{\Gamma\left(\mathrm{Bi}, \gamma^{-1}\bar{\psi}\right)}{\Gamma\left(\mathrm{Ai}, \gamma^{-1}\bar{\psi}\right)}\Gamma\left(\mathrm{Ai}, \gamma^{-1}(1+\bar{\psi})\right)\right| \\ &\qquad < c_1\,\gamma^{1/4}\left[\frac{E(\gamma^{-1}(1+\Psi_L))}{E(\gamma^{-1}\Psi_L)}\right]^2, \end{aligned} \tag{C.2}$$

for some $c_1 > 0$ and γ small enough.

Next, differentiating (C.1), we calculate

$$\begin{aligned} \mathcal{B}'(\gamma^{-1}\bar{\psi}) - \Gamma'(\mathrm{Bi}, \gamma^{-1}(1+\bar{\psi})) &= -\frac{\Gamma\left(\mathrm{Bi}, \gamma^{-1}\bar{\psi}\right)}{\Gamma\left(\mathrm{Ai}, \gamma^{-1}\bar{\psi}\right)}\Gamma\left(\mathrm{Ai}', \gamma^{-1}(1+\bar{\psi})\right) \\ &+ \left(\frac{\Gamma\left(\mathrm{Bi}, \gamma^{-1}\bar{\psi}\right)\Gamma\left(\mathrm{Ai}', \gamma^{-1}\bar{\psi}\right)}{[\Gamma\left(\mathrm{Ai}, \gamma^{-1}\bar{\psi}\right)]^2} - \frac{\Gamma\left(\mathrm{Bi}', \gamma^{-1}\bar{\psi}\right)}{\Gamma\left(\mathrm{Ai}, \gamma^{-1}\bar{\psi}\right)}\right)\Gamma\left(\mathrm{Ai}, \gamma^{-1}(1+\bar{\psi})\right). \end{aligned}$$

Using Lemma A.1 and Corollary A.1 to estimate the right-hand side, we find

$$\begin{aligned} &\sup_{\bar{\psi}\in[\Psi_R,\Psi_L]} \left| E(\gamma^{-1}(1+\bar{\psi}))\left[\mathcal{B}'(\gamma^{-1}\bar{\psi}) - \Gamma'(\mathrm{Bi}, \gamma^{-1}(1+\bar{\psi}))\right]\right| \\ &\qquad < c_1'\,\gamma^{-1/4}\left[\frac{E(\gamma^{-1}(1+\Psi_L))}{E(\gamma^{-1}\Psi_L)}\right]^2, \end{aligned} \tag{C.3}$$

for some $c_1' > 0$ and γ small enough.

The desired result follows from (C.2) and (C.3). □

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Appl. Math. Sci. 35, Springer-Verlag, New York, 1999.

[3] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh, *VODE: A variable-coefficient ODE solver*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1038–1051.

[4] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] U. Ebert, M. Arrayás, N. Temme, B. P. Sommeijer, and J. Huisman, *Critical conditions for phytoplankton blooms*, Bull. Math. Biol., 63 (2001), pp. 1095–1124.

[6] P. G. Falkowski, R. T. Barber, and V. Smetacek, *Biogeochemical controls and feedbacks on ocean primary production*, Science, 281 (1998), pp. 200–206.

[7] K. Fennel and E. Boss, *Subsurface maxima of phytoplankton and chlorophyll: Steady-state solutions from a simple model*, Limnol. Oceanogr., 48 (2003), pp. 1521–1534.

[8] S. Ghosal and S. Mandre, *A simple model illustrating the role of turbulence on phytoplankton blooms*, J. Math. Biol., 46 (2003), pp. 333–346.

[9] M. H. Holmes, *Introduction to Perturbation Methods*, Texts Appl. Math. 20, Springer-Verlag, New York, 1995.

[10] J. Huisman, P. van Oostveen, and F. J. Weissing, *Critical depth and critical turbulence: Two different mechanisms for the development of phytoplankton blooms*, Limnol. Oceanogr., 44 (1999), pp. 1781–1787.

[11] J. Huisman, N. N. Pham Thi, D. M. Karl, and B. P. Sommeijer, *Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum*, Nature, 439 (2006), pp. 322–325.

[12] J. Huisman and B. P. Sommeijer, *Population dynamics of sinking phytoplankton in light-limited environments: Simulation techniques and critical parameters*, J. Sea Res., 48 (2002), pp. 83–96.

[13] W. Hundsdorfer and J. G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Ser. Comput. Math. 33, Springer-Verlag, New York, 2003.

[14] H. Ishii and I. Takagi, *Global stability of stationary solutions to a nonlinear diffusion equation in phytoplankton dynamics*, J. Math. Biol., 16 (1982), pp. 1–24.

[15] C. A. Klausmeier and E. Litchman, *Algal games: The vertical distribution of phytoplankton in poorly mixed water columns*, Limnol. Oceanogr., 46 (2001), pp. 1998–2007.

[16] K. H. Mann and J. R. N. Lazier, *Dynamics of Marine Ecosystems*, Blackwell Science, Oxford, UK, 1996.

[17] N. N. Pham Thi, J. Huisman, and B. P. Sommeijer, *Simulation of three-dimensional phytoplankton dynamics: Competition in light-limited environments*, J. Comput. Appl. Math., 174 (2005), pp. 57–77.

[18] I. Stakgold, *Boundary Value Problems of Mathematical Physics*, Vol. II, Macmillan, New York, 1968.

[19] H. E. de Swart, H. M. Schuttelaars, and S. A. Talke, *Initial growth of phytoplankton in turbid estuaries: A simple model*, Continental Shelf Research, (2007), to appear; doi: 10.1016/j.csr.2007.09.006.

[20] K. Yoshiyama and H. Nakajima, *Catastrophic shifts in vertical distributions of phytoplankton. The existence of a bifurcation set*, J. Math. Biol., 52 (2006), pp. 235–276.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1205–1227

THE COMPLETE CLASSIFICATION FOR DYNAMICS IN A NINE-DIMENSIONAL WEST NILE VIRUS MODEL*

JIFA JIANG† AND ZHIPENG QIU‡

Abstract. Bowman et al. [*Bull. Math. Biol.*, 67 (2005), pp. 1107–1133] proposed a nine-dimensional system of ordinary differential equations modelling West Nile virus in a mosquito–bird-human community and presented some mathematical analysis and its biological explanation. Jiang et al. [*Bull. Math. Biol.*, 2008, DOI 10.1007/s11538-008-9374-6] continued to study the existence and classification of all equilibria and all their local stability and dealt with saddle-node bifurcation of the system. The previous investigation shows that the unique positive equilibrium is globally asymptotically stable if the basic reproduction number is greater than one and the bird death rate is suitably small, but numerical simulations suggest that the unique endemic equilibrium is globally asymptotically stable even if for a large value of the bird death rate. So, they all leave it an open problem. The present paper is to provide a thorough classification of dynamics for this system. In particular, if the reproduction number is greater than one, then a unique endemic equilibrium exists and is globally asymptotically stable in the interior of the feasible region, and the disease persists at an endemic equilibrium if it initially exists, which completely solves the open problem above. Besides, the sufficient and necessary conditions for switch phenomena of the model are obtained if the reproduction number is smaller than one. The results show that the reproduction number alone is not enough to determine whether West Nile virus can prevail or not. Meanwhile, the dynamics of the model for the critical case where the reproduction number is one is also analyzed.

Key words. West Nile virus, differential equations, compound matrices, switch phenomena, global stability

AMS subject classifications. Primary, 92D30; Secondary, 34D23

DOI. 10.1137/070709438

1. Introduction.

Compartmental epidemiological models have played a significant role in the development of a better understanding the mechanism of epidemic transmission and the various preventive strategies used against it. Since the pioneering work of Kermack-Mckendrick, susceptible-infective-recovered (SIR)/susceptible-exposed-infective recovered (SEIR) epidemiological models have received much attention from scientists (see [1, 2, 3] and references therein). West Nile virus (WNV), a single-stranded ribonucleic acid (RNA) virus of the genus *Flavivirus* and the family *Flaviviridae*, remains a significant threat to public health in the world. Since the first outbreak in New York in the late summer of 1999 [4], WNV has been keeping spread through the continent of North America for the last several years. In the United States between 1999 and 2001, WNV was associated with 149 cases of neurological diseases in humans, 814 cases of equine encephalitis, and 11,932 deaths in the avian population. During 2003, 9,858 human cases and 14 deaths were reported [5]. It is, therefore, imperative to gain some insights into the transmission dynamics of WNV

*Received by the editors November 27, 2007; accepted for publication (in revised form) October 2, 2008; published electronically February 4, 2009.

http://www.siam.org/journals/siap/69-5/70943.html

†Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China (jiangjf@shnu.edu.cn). The work of this author was supported by Chinese NSF grants 10671143, 10531030, and Leading Discipline Project of Shanghai Normal University, project DZL707.

‡Department of Applied Mathematics, College of Science, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China (nustqzp@mail.njust.edu.cn). The work of this author was supported by Chinese NSF grant 10801074.

so that we can assess the various anti-WNV preventive strategies. The purpose of the paper is to use mathematical modelling to understand the transmission dynamics of WNV in the mosquito–bird-human population.

A brief review of the salient features of WNV transmission will be useful. WNV is an "arbovirus," which means it is carried by an arthropod, usually an insect, from host to host. WNV is primarily an avian virus and is usually transmitted from bird to bird by mosquitoes. When a mosquito bites an infected bird (to feed on its blood), virus particles in the bird's blood are picked up. These particles make their way to the salivary glands of the mosquito, where they are reproduced. The mosquito then injects them into the next bird it bites. Virus particles injected into susceptible species recognize, and infect, the cells of particular tissues and co-opt those cells into making more virus particles. These particles then leave the cells and circulate in the blood, where they can be picked up by biting mosquitoes; this is how mosquitoes become infected. Humans, horses, and probably other vertebrates are circumstantial hosts; that is, they can be infected if bitten by an infectious mosquito but they do not transmit the disease.

Recently, there has been some effort in the mathematical modelling of the transmission of WNV. Lord and Day [6] carried out simulation studies of St. Louis encephalitis and WNV using a model of differential equations. Thomas and Urena [7] formulated a difference equation model for WNV, targeting its effects on New York City. Wonham, de-Camino-Beck, and Lewis [8] presented a single season ordinary differential equations model for WNV transmission in the mosquito-bird population. Kenkre et al. [9] provided a theoretical framework for the analysis of the WNV epidemic and for dealing with mosquito diffusion and bird migration. Liu et al. [10] studied the impact of the directional dispersal of birds on the spatial spreading of WNV. In a more recent work by Bowman et al. [11], they proposed a single-season ordinary differential equation model for the transmission dynamics of WNV in a mosquito–bird-human community, with birds as reservoir hosts and culicine mosquitoes as vectors. The model in paper [11] is described by the following equations:

$$
\left\{
\begin{aligned}
\frac{dM_u}{dt} &= \Pi_M - \frac{b_1(N_M,N_B,N_H)\beta_1 M_u B_i}{N_B} - \mu_M M_u,\\
\frac{dM_i}{dt} &= \frac{b_1(N_M,N_B,N_H)\beta_1 M_u B_i}{N_B} - \mu_M M_i,\\
\frac{dB_u}{dt} &= \Pi_B - \frac{b_1(N_M,N_B,N_H)\beta_2 M_i B_u}{N_B} - \mu_B B_u,\\
\frac{dB_i}{dt} &= \frac{b_1(N_M,N_B,N_H)\beta_2 M_i B_u}{N_B} - \mu_B B_i - d_B B_i,\\
\frac{dS}{dt} &= \Pi_H - \frac{b_2(N_M,N_B,N_H)\beta_3 M_i S}{N_H} - \mu_H S,\\
\frac{dE}{dt} &= \frac{b_2(N_M,N_B,N_H)\beta_3 M_i S}{N_H} - \mu_H E - \alpha E,\\
\frac{dI}{dt} &= \alpha E - \mu_H I - \delta I,\\
\frac{dH}{dt} &= \delta I - \mu_H H - d_H H - \tau H,\\
\frac{dR}{dt} &= \tau H - \mu_H R.
\end{aligned}
\right.
\tag{1.1}
$$

The above model is based on monitoring the temporal dynamics of the populations of uninfected female mosquitoes $M_u(t)$, infected female mosquitoes $M_i(t)$, uninfected birds $B_u(t)$, infected birds $B_i(t)$, susceptible humans $S(t)$, asymptomatically infected humans $E(t)$, symptomatically infected humans $I(t)$, hospitalized WNV-infected humans $H(t)$, and recovered humans $R(t)$. In (1.1), $N_M(t) = M_u(t) + M_i(t)$ is the total population of female mosquitoes in the community, $N_B(t) = B_u(t) + B_i(t)$ is the total population of birds in the community, and $N_H(t) = S(t) + E(t) + I(t) + H(t) + R(t)$ is the total human population. Π_M, Π_B, and Π_H are the recruitment of uninfected (susceptible) mosquitoes, birds, and human (either by birth or immigration), respectively; $b_1(N_M, N_B, N_H), b_2(N_M, N_B, N_H)$ is the per capita biting rate of mosquitoes on the primary host (birds) and on humans, respectively. In paper [11], it was always assumed that $b_1(N_M, N_B, N_H) = b_1$ and $b_2(N_M, N_B, N_H) = b_2$ are positive constants. β_1, β_2, and β_3 are the probability of WNV transmission from infected birds to uninfected mosquitoes, from mosquitoes to birds, and from mosquitoes to humans, respectively; μ_M, μ_B, and μ_H are the natural death rate of mosquitos, birds, and humans, respectively; d_B, d_H denote the WNV-induced death rate of birds and humans; α is the development rate from asymptomatically infected humans into the symptomatically infected humans; δ is the hospitalization rate from the symptomatic population to the population of hospitalized individuals; τ is the recovery rate from the population of hospitalized individuals into the recovered population.

In paper [11], by investigating the qualitative features of the system, an important epidemiological threshold, known as the basic reproduction number, was determined, and sufficient conditions for the local and global stability of the associated equilibria were obtained; all parameters were estimated by real data, and detailed explanations were given for their theoretic results. Applying the theory of K-competitive dynamical systems [19] and index theory of dynamical systems on a surface, Jiang et al. [12] considered a subsystem for the primary mosquito-bird cycle and obtained sufficient and necessary conditions for local stability of equilibria of the subsystem. The results in paper [12] suggested that the basic reproduction number alone is not enough to determine whether WNV can prevail or not, and that more attention should be paid to the initial state of WNV.

However, there remain some problems. Although the global stability of the endemic equilibrium was proved in [11] with the assumption that d_B is sufficiently small and the basic reproduction number is above unity, Bowman et al. [11] were not able to show the global dynamics of system (1.1) for all $d_B > 0$, and mass numerical simulations in [11, 12] suggest that the unique endemic equilibrium is globally asymptotically stable even for large values of d_B. However, "rigorous proof for the global stability of the endemic equilibrium in the case when the basic reproduction number is greater than 1 remains an open problem" [12]. In addition, paper [12] classified the equilibria and their local stability for the subsystem of system (1.1), which involves only the mosquitoes and birds. Therefore, we expect to obtain the complete dynamical behavior of system (1.1) and give it a good explanation in reality. In this paper we will not only give a positive answer for the above conjecture, but we also completely classify the dynamics of the WNV model even for the critical case.

The remaining part of this paper is organized as follows: In section 2, we mainly classify the existence of equilibria and investigate their local stability. In section 3, we study the dynamics of the limiting subsystem involving only the mosquitoes and birds. Based on the dynamics of the limiting system, we provide a thorough classification for the dynamics of (1.1) in section 4. In the last section we conclude the paper with a discussion.

2. The existence of equilibria and their local stability. Since $N_M(t) = M_u(t) + M_i(t)$, system (1.1) is equivalent to

$$\left\{\begin{array}{l} \frac{dN_M}{dt} = \Pi_M - \mu_M N_M, \\ \frac{dM_i}{dt} = \frac{b_1\beta_1 M_u B_i}{N_B} - \mu_M M_i, \\ \frac{dB_u}{dt} = \Pi_B - \frac{b_1\beta_2 M_i B_u}{N_B} - \mu_B B_u, \\ \frac{dB_i}{dt} = \frac{b_1\beta_2 M_i B_u}{N_B} - \mu_B B_i - d_B B_i, \\ \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_3 M_i S}{N_H} - \mu_H S, \\ \frac{dE}{dt} = \frac{b_2\beta_3 M_i S}{N_H} - \mu_H E - \alpha E, \\ \frac{dI}{dt} = \alpha E - \mu_H I - \delta I, \\ \frac{dH}{dt} = \delta I - \mu_H H - d_H H - \tau H, \\ \frac{dR}{dt} = \tau H - \mu_H R. \end{array}\right. \tag{2.1}$$

Since the equilibrium for N_M is

$$N_M(t) = \frac{\Pi_M}{\mu_M},$$

the limiting system of (2.1) is

$$\left\{\begin{array}{l} \frac{dM_i}{dt} = \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{B_i + B_u} - \mu_M M_i, \\ \frac{dB_u}{dt} = \Pi_B - \frac{b_1\beta_2 M_i B_u}{B_i + B_u} - \mu_B B_u, \\ \frac{dB_i}{dt} = \frac{b_1\beta_2 M_i B_u}{B_i + B_u} - \mu_B B_i - d_B B_i, \\ \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_3 M_i S}{N_H} - \mu_H S, \\ \frac{dE}{dt} = \frac{b_2\beta_3 M_i S}{N_H} - \mu_H E - \alpha E, \\ \frac{dI}{dt} = \alpha E - \mu_H I - \delta I, \\ \frac{dH}{dt} = \delta I - \mu_H H - d_H H - \tau H, \\ \frac{dR}{dt} = \tau H - \mu_H R. \end{array}\right. \tag{2.2}$$

Therefore, the dynamics of (1.1) or (2.1) is qualitatively equivalent to it given by (2.2) (e.g., see [14, 15, 16, 17, 30]), we investigate only (2.2) hereafter. In this section, we mainly investigate the existence of equilibria for system (2.2) and their local stability.

Let

$$\mathcal{D} = \left\{ (M_i, B_u, B_i, S, E, I, H, R) \in \mathbb{R}_+^8 : 0 < M_i \le \frac{\Pi_M}{\mu_M}, \right.$$
$$\left. \frac{\Pi_B}{\mu_B + d_B} < N_B \le \frac{\Pi_B}{\mu_B}, 0 < N_H \le \frac{\Pi_H}{\mu_H} \right\}.$$

Then it follows from paper [11] that all solutions of the system (2.2) starting in $\mathcal{D}$ remain in $\mathcal{D}$ for all $t > 0$. Thus, $\mathcal{D}$ is positively invariant, and it is sufficient to consider solutions in $\mathcal{D}$. In this region, the usual existence, uniqueness, and continuation results hold for the system (2.2). In what follows, we always assume that the initial points lie in $\mathcal{D}$.

It is well known that one of the most important subjects in epidemic models is to obtain a threshold or reproduction number R_0 that determines the persistence and extinction of a disease. There are many papers investigating the reproduction number for systems modelled by ordinary and delay differential equations [1, 2, 3, 18]. The basic reproduction number for the system (2.2) was calculated in [11]:

$$R_0 = \sqrt{\frac{b_1^2 \beta_1 \beta_2 \mu_B \Pi_M}{\mu_M^2 (\mu_B + d_B) \Pi_B}}.$$

Adopting the notations in [12], we denote

$$\begin{aligned} a_2 &= \frac{d_B(\mu_M d_B - b_1 \beta_1 \mu_B)}{b_1 \beta_2 \mu_B}, \\ a_1 &= \frac{b_1 \beta_1 \Pi_M}{\mu_M} - \frac{(2\mu_M d_B - b_1 \beta_1 \mu_B)\Pi_B}{b_1 \beta_2 \mu_B}, \\ a_0 &= \frac{\mu_M \Pi_B^2 (1 - R_0^2)}{b_1 \beta_2 \mu_B}, \\ \Delta &= a_1^2 - 4 a_0 a_2, \\ B_{i2}^* &= \frac{\Pi_B}{\mu_B + d_B}. \end{aligned}$$

Now we are able to state the result on the existence of equilibria for system (2.2).

THEOREM 2.1. *The system* (2.2) *can have up to three equilibria. More precisely, we have the following:*

(1) *The boundary equilibrium, the disease-free equilibrium (DEF)* $E_0(0, \frac{\Pi_B}{\mu_B}, 0, \frac{\Pi_H}{\mu_H}, 0, 0, 0, 0)$ *always exists.*

(2) *Assume* $R_0 > 1$. *Then system* (2.2) *has a unique positive equilibrium* $E^*(M_i^*, B_u^*, B_i^*, S^*, E^*, I^*, H^*, R^*)$, *where* $0 < B_i^* < B_{i2}^* = \frac{\Pi_B}{\mu_B + d_B}$.

(3) *Assume* $R_0 < 1$. *Then we have*

(3a) *if* $a_2 \le 0$, *system* (2.2) *has no positive equilibrium;*

(3b) *if* $a_2 > 0$, *system* (2.2) *has either no or two possible equilibria. Moreover, system* (2.2) *has two positive equilibria* $E^1(M_i^1, B_u^1, B_i^1, S^1, E^1, I^1, H^1, R^1)$ *and* $E^2(M_i^2, B_u^2, B_i^2, S^2, E^2, I^2, H^2, R^2)$, *where* $0 < B_i^1 < B_i^2 < B_{i2}^* = \frac{\Pi_B}{\mu_B + d_B}$, *if and only if*

$$\Delta > 0 \text{ and } 0 < \frac{-a_1}{2a_2} < \frac{\Pi_B}{\mu_B + d_B}.$$

These two equilibria coalesce into one equilibrium $E^1(M_i^1, B_u^1, B_i^1, S^1, E^1, I^1, H^1, R^1)$ *if and only if* $0 < \frac{-a_1}{2a_2} < \frac{\Pi_B}{\mu_B + d_B}$ *and* $\Delta = 0$.

(4) *Assume* $R_0 = 1$. *Then system* (2.2) *has no positive equilibrium if* $\mu_M(d_B - \mu_B) \le b_1 \beta_1 \mu_B$, *and if* $\mu_M(d_B - \mu_B) > b_1 \beta_1 \mu_B$, *system* (2.2) *has a unique positive equilibrium* $E^*(M_i^*, B_u^*, B_i^*, S^*, E^*, I^*, H^*, R^*)$.

The proof of Theorem 2.1 is based on a simple algebraic analysis, and the reader can refer to the papers [11] and [12].

Now let $E^{\#}(M_i^{\#}, B_u^{\#}, B_i^{\#}, S^{\#}, E^{\#}, I^{\#}, H^{\#}, R^{\#})$ be an arbitrary equilibrium of the system (2.2) and

$$N_H^{\#} = S^{\#} + E^{\#} + I^{\#} + H^{\#} + R^{\#}.$$

Then the Jacobian matrix of the vector field corresponding to system (2.2), evaluated at $E^{\#}$, is

$$J(E^{\#}) = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

where

$$A_{11} = \begin{bmatrix} -\left(\mu_M + \dfrac{b_1\beta_1 B_i^{\#}}{B_i^{\#}+B_u^{\#}}\right) & -\dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i^{\#})B_i^{\#}}{(B_i^{\#}+B_u^{\#})^2} & \dfrac{b_1\beta_2 M_i^{\#} B_u^{\#}}{(B_i^{\#}+B_u^{\#})^2} \\ -\dfrac{b_1\beta_2 B_u^{\#}}{B_i^{\#}+B_u^{\#}} & -\left(\dfrac{b_1\beta_2 M_i^{\#} B_i^{\#}}{(B_i^{\#}+B_u^{\#})^2} + \mu_B\right) & \dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i^{\#})B_u^{\#}}{(B_i^{\#}+B_u^{\#})^2} \\ \dfrac{b_1\beta_2 B_u^{\#}}{B_i^{\#}+B_u^{\#}} & \dfrac{b_1\beta_2 M_i^{\#} B_i^{\#}}{(B_i^{\#}+B_u^{\#})^2} & -\left(\mu_B + d_B + \dfrac{b_1\beta_2 M_i^{\#} B_u^{\#}}{(B_i^{\#}+B_u^{\#})^2}\right) \end{bmatrix},$$

$$A_{22} = \left[\begin{array}{ccc} -\dfrac{b_2\beta_3 M_i^{\#}(N_H^{\#} - S^{\#})}{(N_H^{\#})^2} - \mu_H & \dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} & \dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} \\ \dfrac{b_2\beta_3 M_i^{\#}(N_H^{\#} - S^{\#})}{(N_H^{\#})^2} & -\dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} - \mu_H - \alpha & -\dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} \\ 0 & \alpha & -(\mu_H + \delta) \\ 0 & 0 & \delta \\ 0 & 0 & 0 \end{array}\right.$$

$$\left.\begin{array}{cc} \dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} & \dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} \\ -\dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} & -\dfrac{b_2\beta_3 M_i^{\#} S^{\#}}{(N_H^{\#})^2} \\ 0 & 0 \\ -(\mu_H + d_H + \tau) & 0 \\ \tau & -\mu_H \end{array}\right],$$

and

$$A_{21} = \begin{bmatrix} -\dfrac{b_2\beta_3 S^{\#}}{N_H^{\#}} & 0 & 0 \\ \dfrac{b_2\beta_3 S^{\#}}{N_H^{\#}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

After extensive algebraic calculations [11], the characteristic equation associated with A_{22} is given by

$$(\lambda+\mu_H)\Bigg[(\lambda+\mu_H+d_H+\tau)(\lambda+\mu_H+\delta)(\lambda+\mu_H+\alpha)\left(\lambda+\mu_H+\frac{b_2\beta_3 M_i^{\#}}{N_H^{\#}}\right)-\frac{d_H\delta\alpha b_2\beta_3 M_i^{\#}S^{\#}}{(N_H^{\#})^2}\Bigg]=0.$$

The constant term in the above polynomial is positive, and the Routh–Hurwitz conditions show that all eigenvalues of the matrix A_{22} have negative real parts. Thus the stability of the equilibrium $E^{\#}$ is determined by the eigenvalues of the matrix A_{11}, since the eigenvalues of the matrix $J(E^{\#})$ are made up of the eigenvalues of the matrices A_{11} and A_{22}. The stability of the matrix A_{11} is discussed in paper [12]. Combining Theorems 3.4 and 3.5 in paper [12] and the stability of the matrix A_{22}, we have the following results.

THEOREM 2.2. (1) *Assume $R_0 > 1$. Then the disease-free equilibrium E_0 is unstable, and the unique positive equilibrium E^* is locally asymptotically stable.*

(2) *Assume $R_0 < 1$. Then the unique boundary equilibrium E_0 is locally asymptotically stable, and we have*

(2a) *if $a_2 > 0$, $\Delta > 0$, and $0 < \frac{-a_1}{2a_2} < \frac{\Pi_B}{\mu_B+d_B}$, then the positive equilibrium E^1 is a saddle point and the positive equilibrium E^2 is locally asymptotically stable. Moreover,* $\dim W^s(E^1)=7$, $\dim W^u(E^1)=1$, *and* $\dim W^s(E^2)=8$*;*

(2b) *if $a_2 > 0, 0 < \frac{-a_1}{2a_2} < \frac{\Pi_B}{\mu_B+d_B}$, and $\Delta = 0$, then the unique positive equilibria E^1 is unstable and* $\dim W^s(E^1)=7$, $\dim W^c(E^1)=1$.

3. The dynamics of a subsystem.

Since humans do not feed back into the mosquito-bird cycle, the subsystem involving only the mosquitoes and birds is independent. The dynamics of the subsystem for the primary mosquito-bird cycle is governed by the following three-dimensional differential equations:

$$\begin{cases} \frac{dM_i}{dt} = \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M}-M_i)B_i}{B_i+B_u} - \mu_M M_i,\\ \frac{dB_u}{dt} = \Pi_B - \frac{b_1\beta_2 M_i B_u}{B_i+B_u} - \mu_B B_u,\\ \frac{dB_i}{dt} = \frac{b_1\beta_2 M_i B_u}{B_i+B_u} - \mu_B B_i - d_B B_i. \end{cases} \tag{3.1}$$

In this section, we mainly provide a complete classification for system (3.1). In paper [12], Jiang et al. provided the classification for equilibria of system (3.1) and by applying the theory of K-competitive dynamical systems [19] and index theory of dynamical systems on a surface, sufficient and necessary conditions for local stability of equilibria are also obtained if the basic reproduction number is not unity. In the following, we will study the dynamics of system (3.1) further and provide a through classification of dynamics for system (3.1).

First, let us present some preliminary results for system (3.1). As in [12], we set

$$\Gamma = \left\{(M_i,B_u,B_i)|\ 0\le M_i\le \frac{\Pi_M}{\mu_M}, \frac{\Pi_B}{\mu_B+d_B}\le B_u+B_i\le\frac{\Pi_B}{\mu_B}\right\}.$$

Then we have what follows.

PROPOSITION 3.1 (see [12]). *All solutions of the system* (3.1) *with nonnegative initial conditions remain nonnegative. Moreover, Γ is a global attractor in $\mathbb{R}^3_+$ and positively invariant for* (3.1).

In what follows, we always assume that the initial points $(M_i(0), B_u(0), B_i(0))$ of system (3.1) lie in Γ. From the last section, it is easy to verify that if $E^{\#}(M_i^{\#}, B_u^{\#}, B_i^{\#},$

$S^{\#}, E^{\#}, I^{\#}, H^{\#}, R^{\#})$ is an equilibrium of (2.2), then $\hat{E}^{\#}(M_i^{\#}, B_u^{\#}, B_i^{\#})$ is an equilibrium of (3.1), and $E^{\#}$ and $\hat{E}^{\#}$ have the same stability. For convenience, we still denote the equilibrium $\hat{E}^{\#}$ by $E^{\#}$.

By analyzing the vector field of system (3.1) on the boundary $\partial\mathbb{R}^3_+ \bigcap \Gamma$, we have that $\partial\mathbb{R}^3_+ \bigcap \Gamma$ contains only positive B_u-axis as its invariant set, in other words, any positive orbit from a point in $\partial\mathbb{R}^3_+ \bigcap \Gamma$ but not in the positive B_u-axis will enter $\text{Int}\mathbb{R}^3_+$. In particular, any periodic orbit if it exists lies in $\text{Int}\mathbb{R}^3_+$.

Let $M = Df(E_0)$ and $\lambda_1 < \lambda_2 = -\mu_b < \lambda_3$ be its eigenvalues and $v_1, v_2 = (0, 1, 0), v_3$ be their eigenvector, respectively. From the Perron–Frobenius theorem, we may assume $v_1 \gg_K 0$. Let Π be the plane spanned by v_1 and v_2. We claim that if any solution $(M_u(t), B_u(t), B_i(t))$ is convergent to E_0 as $t \to \infty$ and tangent to Π at E_0, then $M_u(t) = B_i(t) = 0$. This implies that there is neither homoclinic nor heteroclinic orbit initiating from E_0.

PROPOSITION 3.2. (1) *If the system* (3.1) *has a periodic orbit, then the periodic orbit lies in* $\text{Int}\mathbb{R}^3_+$*;*

(2) *The system* (3.1) *has neither homoclinic nor heteroclinic orbit from* E_0.

The Jacobian of system (3.1) is

$$\begin{bmatrix} -\left(\mu_M + \dfrac{b_1\beta_1 B_i}{B_i + B_u}\right) & -\dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} & \dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} \\ -\dfrac{b_1\beta_2 B_u}{B_i + B_u} & -\left(\dfrac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} + \mu_B\right) & \dfrac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} \\ \dfrac{b_1\beta_2 B_u}{B_i + B_u} & \dfrac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} & -\left(\mu_B + d_B + \dfrac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) \end{bmatrix}.$$

It follows from paper [19] that the system (3.1) is K-competitive in Γ, with $K = \{(M_i, B_u, B_i) \mid M_i > 0, B_u > 0, B_i < 0\}$. From the expressions of $M_i^1, B_u^1, B_i^1, M_i^2, B_u^2, B_i^2, M_i^*, B_u^*, B_i^*$, it is easy to see that the equilibria E^1, E^2, E_0 or E_0, E^* are unordered in the K-order. It follows from Proposition 3.2 in [20] and Proposition 1.3 in [21] that there exists a two-dimensional compact Lipschitz submanifold Σ such that $E^1, E^2 \in \text{Int}\Sigma$ or $E^* \in \text{Int}\Sigma$, $E_0 \in \partial\Sigma$. Moreover, Σ is K-balanced. Since Σ is a two-dimensional compact Lipschitz submanifold and homeomorphic to a compact domain in the plane, it is obvious that the Poincaré–Bendixson theorem holds for the dynamics of (3.1) on Σ. This finding is the key point we can completely analyze for the global behavior for the system (3.1). Moreover, we have the following useful result.

PROPOSITION 3.3. *If the system* (3.1) *has no positive equilibrium, then the disease-free equilibrium* E_0 *is globally asymptotically stable in* Γ.

Proof. Since system (3.1) has the Pioncaré–Bendixson property, the assumption that system (3.1) has no positive equilibrium implies that there is no nontrivial periodic orbit for (3.1). Thus every positive limit set contains the unique equilibrium E_0. If E_0 is not globally asymptotically stable, then there is an entire orbit converging to E_0 in either direction, since asymptotical behavior for three-dimensional competitive system is reduced to two dimension (see [19]), i.e., system (3.1) has a homoclinic orbit from E_0. This contradicts Proposition 3.2. The contradiction implies that the disease-free equilibrium E_0 of system (3.1) is globally asymptotically stable and completes the proof of Proposition 3.3 □

3.1. The global stability of system (3.1) for the case $R_0 > 1$. In this subsection we mainly study the dynamics of system (3.1) for the case $R_0 > 1$. We start with a general mathematical framework for proving global stability, which will be used to prove the principal result in this subsection. The framework is developed in the paper of Muldowney [23, 24, 27]. The presentation here follows from [24].

Let $x \to f(x) \in \mathbb{R}^3$ be a C^1 function defined in $\mathbb{R}^3_+$. We consider the autonomous system in $\Gamma \subset \mathbb{R}^3_+$

$$\dot{x} = f(x). \tag{3.2}$$

Let $x(t, x_0)$ denote the solution of (3.2) such that $x(0, x_0) = x_0$. The linear variational equation of (3.2) with respect to $x(t, x_0)$ is given by

$$\dot{y}(t) = Df(x(t, x_0))y(t), \tag{3.3}$$

where Df is the Jacobian matrix of f. *The second compound equation* with respect to the solution $x(t, x_0) \in \Gamma$ to (3.3) can be described by

$$\dot{z}(t) = Df^{[2]}(x(t, x_0))z(t). \tag{3.4}$$

$Df^{[2]}$ is the second additive compound matrix of the Jacobian matrix Df of f. Generally speaking, for a 3×3 matrix $A = (a_{ij})_{3\times 3}$, *the second additive compound matrix* of A is the matrix $A^{[2]}$ defined as follows:

$$\begin{bmatrix} a_{11} + a_{22} & a_{23} & -a_{13} \\ a_{32} & a_{11} + a_{33} & a_{12} \\ -a_{31} & a_{21} & a_{22} + a_{33} \end{bmatrix}.$$

Suppose system (3.2) has a periodic solution $x = p(t)$, with the least positive period $\omega > 0$ and the orbit $\gamma = \{p(t) : 0 \le t \le \omega\}$. The orbit is *orbitally stable* if for each $\epsilon > 0$, there exists a $\delta > 0$ such that any solution $x(t)$, for which the distance of $x(0)$ from γ is less than δ, remains at a distance less than ϵ from γ for all $t \ge 0$. It is *asymptotically orbitally stable* if the distance of $x(t)$ from γ also tends to zero as $t \to \infty$. This orbit γ is asymptotically orbitally stable with asymptotic phase if it is asymptotically orbitally stable and there is a $b > 0$ such that any solution $x(t)$, for which the distance of $x(0)$ from γ is less than b, satisfies $|x(t) - p(t - \nu)| \to 0$ as $t \to \infty$ for some ν which may depend on $x(0)$.

The following is a criterion given in [23] and [24].

Lemma 3.4 (see [24]). *A sufficient condition for a periodic orbit $\gamma = \{p(t) : 0 \le t \le \omega\}$ of* (3.2) *to be asymptotically orbitally stable with asymptotic phase is that the linear system*

$$\dot{z}(t) = Df^{[2]}(p(t))z(t) \tag{3.5}$$

is asymptotically stable.

Theorem 3.5. *The trajectory of any nonconstant periodic solution to* (3.1), *if it exists, is asymptotically orbitally stable with asymptotic phase.*

Proof. The Jacobian matrix $J(M_i, B_u, B_i)$ of system (3.1) is given by

$$\begin{bmatrix} -\left(\mu_M + \dfrac{b_1\beta_1 B_i}{B_i + B_u}\right) & -\dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} & \dfrac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} \\ -\dfrac{b_1\beta_2 B_u}{B_i + B_u} & -\left(\dfrac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} + \mu_B\right) & \dfrac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} \\ \dfrac{b_1\beta_2 B_u}{B_i + B_u} & \dfrac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} & -\left(\mu_B + d_B + \dfrac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) \end{bmatrix}.$$

Suppose that the solution $(M_i(t), B_u(t), B_i(t))$ is periodic with the least period $\omega > 0$. The second compound system of (3.1) along a periodic solution $(M_i(t), B_u(t), B_i(t))$ is

$$\begin{aligned} \dot{X} &= -\left(\mu_M + \mu_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2}\right) X + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} Y \\ &\quad - \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} Z, \\ \dot{Y} &= \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} X - \left(\mu_M + \mu_B + d_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) Y \\ &\quad - \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} Z, \\ \dot{Z} &= -\frac{b_1\beta_2 B_u}{B_i + B_u} X - \frac{b_1\beta_2 B_u}{B_i + B_u} Y - \left(2\mu_B + d_B + \frac{b_1\beta_2 M_i}{B_i + B_u}\right) Z. \end{aligned} \tag{3.6}$$

To show that (3.6) is asymptotically stable, we construct a Lyapunov function

$$V(X, Y, Z; M_i, B_u, B_i) = \sup\left\{\frac{B_i}{M_i}(|X| + |Y|), |Z|\right\}.$$

It follows from Proposition 3.2 that any periodic orbit, if it exists, lies in $\text{Int}\mathbb{R}^3_+$. Thus any periodic solution $(M_i(t), B_u(t), B_i(t))$ is at a positive distance from the boundary, and there exists a constant $c_1 > 0$ such that

$$V(X, Y, Z; M_i, B_u, B_i) \geq c_1 \sup\{|X|, |Y|, |Z|\}$$

for all $(X, Y, Z) \in \mathbb{R}^3$ and $(M_i(t), B_u(t), B_i(t)), t \in [0, \omega]$. The right derivative of V along a solution $(X(t), Y(t), Z(t))$ to (3.6) and $(M_i(t), B_u(t), B_i(t))$ can be estimated as follows:

(3.7)

$$\begin{aligned} D_+|X(t)| &\leq -\left(\mu_M + \mu_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2}\right) |X(t)| \\ &\quad + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} |Y(t)| + \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} |Z(t)|, \\ D_+|Y(t)| &\leq \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} |X(t)| - \left(\mu_M + \mu_B + d_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) |Y(t)| \\ &\quad + \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} |Z(t)| \end{aligned}$$

and

(3.8)

$$\begin{aligned} D_+|Z(t)| &\leq \frac{b_1\beta_2 B_u}{B_i + B_u} |X(t)| + \frac{b_1\beta_2 B_u}{B_i + B_u} |Y(t)| - \left(2\mu_B + d_B + \frac{b_1\beta_2 M_i}{B_i + B_u}\right) |Z(t)| \\ &= \frac{b_1\beta_2 B_u M_i}{(B_i + B_u)B_i} \frac{B_i}{M_i}(|X(t)| + |Y(t)|) - \left(2\mu_B + d_B + \frac{b_1\beta_2 M_i}{B_i + B_u}\right) |Z(t)|. \end{aligned}$$

Therefore,

$$\begin{aligned}
&D_+\frac{B_i}{M_i}(|X(t)|+|Y(t)|)\\
&=\left(\frac{\dot{B}_i}{B_i}-\frac{\dot{M}_i}{M_i}\right)\frac{B_i}{M_i}(|X(t)|+|Y(t)|)+\frac{B_i}{M_i}D_+(|X(t)|+|Y(t)|)\\
&\le\left(\frac{\dot{B}_i}{B_i}-\frac{\dot{M}_i}{M_i}-\mu_M-\mu_B-\frac{b_1\beta_1B_i}{B_i+B_u}\right)\frac{B_i}{M_i}(|X(t)|+|Y(t)|)\\
&\quad+\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M}-M_i)B_i}{(B_i+B_u)M_i}|Z(t)|.
\end{aligned}$$

We conclude that (3.7) and (3.8) lead to

$$D_+V(t)\le\sup\{g_1(t),g_2(t)\}V(t), \tag{3.9}$$

where

$$g_1(t)=\frac{\dot{B}_i}{B_i}-\frac{\dot{M}_i}{M_i}-\mu_M-\mu_B-\frac{b_1\beta_1B_i}{B_i+B_u}+\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M}-M_i)B_i}{(B_i+B_u)M_i}, \tag{3.10}$$

$$g_2(t)=\frac{b_1\beta_2B_uM_i}{(B_i+B_u)B_i}-2\mu_B-d_B-\frac{b_1\beta_2M_i}{B_i+B_u}. \tag{3.11}$$

Rewriting (3.1), we find that

$$\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M}-M_i)B_i}{(B_i+B_u)M_i}=\frac{\dot{M}_i}{M_i}+\mu_M,\frac{b_1\beta_2B_uM_i}{(B_i+B_u)B_i}=\frac{\dot{B}_i}{B_i}+\mu_B+d_B. \tag{3.12}$$

From (3.10)–(3.12) we have

$$\sup\{g_1(t),g_2(t)\}\le\frac{\dot{B}_i}{B_i}-\mu_B,$$

and thus

$$\int_0^\omega\sup\{g_1(t),g_2(t)\}dt\le\log B_i(t)|_0^\omega-\mu_B\omega=-\mu_B\omega<0,$$

since $B_i(t)$ is periodic with the least period ω. This relation and (3.9) imply that $V(t)\to0$ as $t\to+\infty$, and, in turn, that $(X(t),Y(t),Z(t))\to0$ as $t\to+\infty$. As a result, the linear system (3.6) is asymptotically stable, and the periodic solution $(M_i(t),B_u(t),B_i(t))$ is asymptotically orbitally stable with asymptotic phase by Lemma 3.4. This completes the proof. ▯

Our principal result in this subsection can be stated as follows.

THEOREM 3.6. *Assume that* $R_0>1$. *Then the unique endemic equilibrium* E^* *is globally stable in* IntΓ.

Proof. Firstly, we claim that system (3.1) has no nontrivial periodic orbit if $R_0>1$. Suppose not, then there is a nontrivial periodic orbit γ with every point positive. Therefore, there exists a two-dimensional compact Lipschitz submanifold Σ, which must contain E^* in its interior. Without loss of generality, we may assume that γ is the nearest periodic orbit from E^* in Σ. From the stability of E^*, γ is unstable on Σ, contradicting Theorem 3.5. This implies that every positive limit set contains an equilibrium.

Secondly, we analyze the geometrical behavior of the stable manifold $W^s(E_0)$. Direct calculation yields that $\lambda_1 < \lambda_2 = -\mu_B < 0 < \lambda_3$, where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of the linearized matrix at E_0. Therefore, $W^s(E_0)$ is tangent to the plane Π spanned by v_1 and v_2. Therefore, $W^s(E_0) \bigcap \mathbb{R}^3_+ = \{(0, \mu, 0) : \mu > 0\}$. This shows that any solution from a positive initial point cannot converge to E_0.

Finally, for any initial point $P = (M_i(0), B_u(0), B_i(0))$, with $(M_i(0))^2 + (B_i(0))^2 \neq 0$, we conclude that $\omega(P) = \{E^*\}$. Otherwise, there exists such a P such that $\omega(P) \neq \{E^*\}$. By the stability of E^*, E^* is not in $\omega(P)$. By the arguments in the first and second paragraph, $\omega(P)$ contains E_0 and is not a singleton. Since asymptotical behavior for the three-dimensional competitive system is reduced to two dimensions (see [19]), $\omega(P)$ contains an entire orbit which is convergent to E_0 as $t \to \pm\infty$. The second paragraph implies that such an entire orbit lies in positive B_u-axis, a contradiction. This completes the proof of Theorem 3.6. □

We remark that Theorem 3.5 also holds for the case $R_0 \leq 1$. In the proof of Theorem 3.6, we use only Theorem 3.5 and the fact that the unique positive equilibrium E^* is locally asymptotically stable and the boundary equilibrium E_0 is unstable, with the stable manifold two-dimensional and unstable manifold one-dimensional.

3.2. The global stability of system (3.1) for the critical case $R_0 = 1$. Now let us state the main result of this subsection.

THEOREM 3.7. *Assume $R_0 = 1$. If $\mu_M(d_B - \mu_B) \leq b_1\beta_1\mu_B$, then the disease-free equilibrium E_0 of system* (3.1) *is globally asymptotically stable; if $\mu_M(d_B - \mu_B) > b_1\beta_1\mu_B$, the unique positive equilibrium E^* of system* (3.1) *is globally asymptotically stable in* $\mathrm{Int}\mathbb{R}^3_+$.

Proof. If $R_0 = 1$ and $\mu_M(d_B - \mu_B) \leq b_1\beta_1\mu_B$, it follows from Theorem 2.1 that the disease-free equilibrium E_0 is the unique equilibrium. Proposition 3.3 implies that the disease-free equilibrium E_0 of system (3.1) is globally asymptotically stable and completes the proof of the first conclusion.

Now we present the proof of the second conclusion. Assume that $R_0 = 1$ and $\mu_M(d_B - \mu_B) > b_1\beta_1\mu_B$. It follows from Theorem 2.1 that system (3.1) has a unique boundary equilibrium E_0 and a unique positive equilibrium E^*. Recall the remark for Theorem 3.6, in order to prove the second conclusion, it then suffices to show that E^* is locally asymptotically stable and E_0 is unstable, with the stable manifold two-dimensional and unstable manifold one-dimensional.

We first claim that E^* is locally asymptotically stable. After extensive algebraic calculations [12], the characteristic equation of the linearized system of (3.1) at E^* can be read

$$\lambda^3 + A_1\lambda^2 + A_2\lambda + A_3 = 0, \tag{3.13}$$

where

$$\begin{aligned}
A_1 = {} & \frac{1}{(\Pi_B - d_B B_i^*)(\Pi_B - (\mu_B + d_B)B_i^*)}[b_1\beta_1\mu_B B_i^*(\Pi_B - (\mu_B + d_B)B_i^*) \\
& + (\Pi_B - d_B B_i^*)[(\mu_M + 2\mu_B + d_B)(\Pi_B - (\mu_B + d_B)B_i^*) + (\mu_B + d_B)\mu_B B_i^*]]; \\
A_2 = {} & \frac{\mu_M\mu_B\Pi_B}{(\Pi_B - (\mu_B + d_B)B_i^*)} + \frac{(\mu_B + d_B)\mu_B[(\Pi_B - d_B B_i^*)^2 + d_B\mu_B B_i^* B_i^*]}{(\Pi_B - d_B B_i^*)(\Pi_B - (\mu_B + d_B)B_i^*)} \\
& + \frac{\Pi_B\mu_B^2 b_1\beta_1 B_i^*}{(\Pi_B - d_B B_i^*)(\Pi_B - (\mu_B + d_B)B_i^*)} + \frac{b_1\beta_1\mu_B(\mu_B + d_B)B_i^*}{\Pi_B - d_B B_i^*}; \\
A_3 = {} & \frac{(\mu_B + d_B)\mu_B B_i^*}{(\Pi_B - d_B B_i^*)^2(\Pi_B - (\mu_B + d_B)B_i^*)}[d_B(\mu_B + d_B)(b_1\beta_1\mu_B - d_B\mu_M)(B_i^*)^2 \\
& + 2\Pi_B d_B(d_B\mu_M - b_1\beta_1\mu_B)B_i^* + \Pi_B^2(b_1\beta_1\mu_B + \mu_M\mu_B - d_B\mu_M)].
\end{aligned}$$

We also have

$$A_1A_2 - A_3 = D_2b_1^2 + D_1b_1 + D_0,$$

where

$$\begin{aligned}
D_2 &= \frac{((\mu_B+d_B)(\Pi_B-(d_B+\mu_B)B_i^*)+\Pi_B\mu_B)(\beta_1\mu_BB_i^*)^2}{(\Pi_B-d_BB_i^*)^2(\Pi_B-(\mu_B+d_B)B_i^*)};\\
D_1 &= \frac{\mu_B\beta_1B_i^*\mu_M((d_B+\mu_B)(\Pi_B-(d_B+\mu_B)B_i^*)+2\Pi_B\mu_B)}{(\Pi_B-d_BB_i^*)(\Pi_B-(\mu_B+d_B)B_i^*)}\\
&\quad + \frac{\mu_B\beta_1B_i^*((d_B+\mu_B)(\Pi_B-(d_B+\mu_B)B_i^*)+\Pi_B\mu_B)^2}{(\Pi_B-d_BB_i^*)(\Pi_B-(\mu_B+d_B)B_i^*)^2};\\
D_0 &= \frac{\Pi_B\mu_B\mu_M^2}{\Pi_B-(\mu_B+d_B)B_i^*} + \frac{1}{(\Pi_B-d_BB_i^*)(\Pi_B-(\mu_B+d_B)B_i^*)^2}\\
&\quad \times\{\Pi_B\mu_B\mu_M((\mu_B+d_B)(\Pi_B-(\mu_B+d_B)B_i^*)^2\\
&\quad +(2\mu_B+d_B)(\Pi_B-(\mu_B+d_B)B_i^*)(\Pi_B-d_BB_i^*)+(\mu_B+d_B)\mu_BB_i^*(\Pi_B-d_BB_i^*))\\
&\quad +\mu_B(\mu_B+d_B)[(d_B+\mu_B)(\Pi_B-(\mu_B+d_B)B_i^*)\\
&\quad +\Pi_B\mu_B]((\Pi_B-d_BB_i^*)^2+\mu_Bd_B(B_i^*)^2)\}.
\end{aligned}$$

Thus it is clear that

$$A_1A_2 - A_3 = D_2b_1^2 + D_1b_1 + D_0 > 0, A_1 > 0,$$

since $0 < B_i^* < \frac{\Pi_B}{\mu_B+d_B}$.

Let $\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3$ be the roots of (3.13) and assume $\mathrm{Re}\bar{\lambda}_1 \le \mathrm{Re}\bar{\lambda}_2 \le \mathrm{Re}\bar{\lambda}_3$. The Perron–Frobenius theorem implies that $\bar{\lambda}_1 < 0$. Since $R_0 = 1$ and $\mu_M(d_B - \mu_B) > b_1\beta_1\mu_B$, direct calculation yields that $\Delta \neq 0$. It follows from Proposition 3.2 in paper [12] that $A_3 \neq 0$. The compact of Σ, together with the fact that the dynamics of system (3.1) on Σ is positively invariant, implies that $A_3 = -\bar{\lambda}_1\bar{\lambda}_2\bar{\lambda}_3 > 0$. Thus we have $A_1 > 0, A_3 > 0, A_1A_2 - A_3 > 0$. By the Routh–Huriwitz criterion, we obtain that $\mathrm{Re}\bar{\lambda}_1 < 0, \mathrm{Re}\bar{\lambda}_2 < 0, \mathrm{Re}\bar{\lambda}_3 < 0$. Thus the positive equilibrium E^* is locally asymptotically stable.

It is easy to see that the linearized matrix for (3.1) at E_0 has two negative eigenvalues and 0. Suppose that E_0 is locally asymptotically stable. Then, pick up a point which is neither in the basin of attraction for E^* nor in the basin of attraction for E_0. The positive limit set for this point contains no equilibrium and hence is a nontrivial periodic orbit, contradicting Theorem 3.5. Thus, E_0 is unstable with the stable manifold two-dimensional and unstable manifold one-dimensional.

Using the same way as in the proof of Theorem 3.6, we conclude that E^* is globally asymptotically stable in this case. □

3.3. The global stability of system (3.1) for the case $R_0 < 1$.

In this subsection, we mainly prove the following principal results.

THEOREM 3.8. *Assume that $R_0 < 1$. The dynamics of system (3.1) are determined by the number of the positive equilibria, and there are three cases:*

(1) *If $a_2 > 0$, $0 < \frac{-a_1}{2a_2} < B_{i2}^*$, and $\Delta > 0$, then system (3.1) has switch phenomenon. That is, a two-dimensional stable manifold for E^1 separates $\mathrm{int}\mathbb{R}_+^3$ into two parts, denoted by V, U, respectively. Part V contains E_0 and one branch $W_1^u(E^1)$ of the unstable manifold for E^1, and part U contains E^2 and the other branch $W_2^u(E^1)$ of the unstable manifold for E^1. Moreover, if $(M_i(0), B_u(0), B_i(0)) \in V$,*

then $(M_i(t), B_u(t), B_i(t)) \to E_0$ *as* $t \to \infty$*; if* $(M_i(0), B_u(0), B_i(0)) \in U$*, then* $(M_i(t), B_u(t), B_i(t)) \to E^2$ *as* $t \to \infty$*; otherwise,* $(M_i(0), B_u(0), B_i(0)) \in W^s(E^1)$ *and* $(M_i(t), B_u(t), B_i(t)) \to E^1$ *as* $t \to \infty$*.*

(2) *If* $a_2 > 0$*,* $0 < \frac{-a_1}{2a_2} < B^*_{i2}$*, and* $\Delta = 0$*, then system* (3.1) *also has a two-dimensional stable manifold for* E^1 *which also separates* $\mathrm{int}\mathbb{R}^3_+$ *into two parts, denoted by* V, U*, respectively. Part* V *contains* E_0 *and one branch* $W^c_1(E^1)$ *of the center manifold for* E^1*, and part* U *contains no positive equilibrium and the other branch* $W^c_2(E^1)$ *of the center manifold for* E^1*. Moreover, if* $(M_i(0), B_u(0), B_i(0)) \in V$*, then* $(M_i(t), B_u(t), B_i(t)) \to E_0$ *as* $t \to \infty$*; otherwise,* $(M_i(t), B_u(t), B_i(t)) \to E^1$ *as* $t \to \infty$*.*

(3) *Otherwise, the disease-free equilibrium* E_0 *is globally asymptotically stable.*

Proof. First, let us consider the case $a_2 > 0$, $0 < \frac{-a_1}{2a_2} < B^*_{i2}$, and $\Delta > 0$. By Theorems 2.1 and 2.2, (3.1) has three equilibria E_0, E^1, E^2. Moreover, E_0, E^2 are locally asymptotically stable, the positive equilibrium E^1 is a saddle point, and $W^s(E^1) = 2, W^u(E^1) = 1$. Let U be the basin of attraction of the positive equilibrium E^2 in $\mathrm{Int}\mathbb{R}^3_+$. Then U is an open *simply connected set*, i.e., each closed curve in U can be continuously deformed to a point within U. Let $\overline{U}$ be the closure of the set U in $\mathrm{Int}\mathbb{R}^3_+$. Since E_0 is locally asymptotically stable, it follows that $E_0 \notin \overline{U}$. The result that any positive orbit from a point in $\partial\mathbb{R}^3_+$ but not in positive B_u-axis will enter $\mathrm{Int}\mathbb{R}^3_+$ implies that there exists constant $\xi > 0$ such that

$$\liminf_{t\to\infty} M_i(t) > \xi, \liminf_{t\to\infty} B_u(t) > \xi, \liminf_{t\to\infty} B_i(t) > \xi,$$

provided that the initial points $(M_i(0), B_u(0), B_i(0))$ lie in $\overline{U}$ by [25]. Thus there exists a compact absorbing set $J \subset \overline{U}$.

Denote by W the Euclidean unit ball in R^2, and let $\overline{W}$ and ∂W be its closure and boundary, respectively. A function $\varphi \in Lip(\overline{W} \to \overline{U})$ will be described as a simply connected rectifiable 2-surface in $\overline{U}$; a function $\psi \in Lip(\partial W \to \overline{U})$ is a closed rectifiable curve in $\overline{U}$ and will be called simple if it is one to one.

We claim that there is not any simple closed rectifiable curves which are invariant with respect to system (3.1) in $\overline{U}$. Assume that our claim is not true, i.e., there exists a simple closed rectifiable curve $\psi \in Lip(\partial W \to \overline{U})$ in $\overline{U}$, which is invariant with respect to system (3.1). Since system (3.1) is K-competitive and $\mathrm{dimW^u(E^1)} = 1$, we have that $\psi \in \Sigma$.

Denote

$$\Pi(\psi, \overline{U}) = \{\varphi \in Lip(\overline{W} \to \overline{U}) : \varphi(\partial W) = \psi(\partial W)\}.$$

Then $\Pi(\psi, \overline{U})$ is nonempty, since Σ is a two-dimensional compact Lipschitz manifold. Define a functional S on $\Pi(\psi, \overline{U})$ by

$$S\varphi = \int_{\overline{W}} \left| P(\varphi) \frac{\partial \varphi}{\partial u_1} \wedge \frac{\partial \varphi}{\partial u_2} \right|,$$

where $(u_1, u_2) \in \overline{W}$, $\wedge$ is Grassman product, and

$$P(M_i, B_u, B_i) = \begin{pmatrix} \frac{B_i}{M_i} & 0 & 0 \\ 0 & \frac{B_i}{M_i} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

From Proposition 2.2 of [26] and the fact that $|P^{-1}(M_i, B_u, B_i)|$ is uniformly bounded for (M_i, B_u, B_i) in any compact subset of $\overline{U}$, for each compact $F \subset \overline{U}$ there exists

$\delta > 0$ such that

$$S\varphi \geq \delta \tag{3.14}$$

for all $\varphi \in \Pi(\psi, \overline{U})$ such that $\varphi(\overline{W}) \subset F$. Let $x = (M_i, B_u, B_i)$ and $f(x)$ denote the vector field of (3.1), and let $\varphi_t = x(t, \varphi)$. Then $y_i(t) = \frac{\partial \varphi_t}{\partial u_i}, i = 1, 2$, are solutions of the linear variational equation of (3.1)

$$\dot{y}(t) = Df(x(t, \varphi))y(t), \tag{3.15}$$

where

$$Df(x(t,\varphi)) = \begin{bmatrix} -(\mu_M + \frac{b_1\beta_1 B_i}{B_i + B_u}) & -\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} & \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} \\ -\frac{b_1\beta_2 B_u}{B_i + B_u} & -\left(\frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} + \mu_B\right) & \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} \\ \frac{b_1\beta_2 B_u}{B_i + B_u} & \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} & -\left(\mu_B + d_B + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) \end{bmatrix}$$

and $z(t) = \frac{\partial \varphi_t}{\partial u_1} \wedge \frac{\partial \varphi_t}{\partial u_2}$ is a solution of the second compound equation of (3.15) (see [22, 23])

$$\dot{z}(t) = Df^{[2]}(x(t, \varphi))z(t), \tag{3.16}$$

where

$$Df^{[2]}(x(t,\varphi)) = \begin{bmatrix} -\left(\mu_M + \mu_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2}\right) & \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2} & -\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i + B_u)^2} \\ \frac{b_1\beta_2 M_i B_i}{(B_i + B_u)^2} & -\left(\mu_M + \mu_B + d_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_u}{(B_i + B_u)^2}\right) & -\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)^2} \\ -\frac{b_1\beta_2 B_u}{B_i + B_u} & -\frac{b_1\beta_2 B_u}{B_i + B_u} & -\left(2\mu_B + d_B + \frac{b_1\beta_2 M_i}{B_i + B_u}\right) \end{bmatrix}. \tag{3.17}$$

Straightforward differentiation shows that $w(t) = P(\varphi_t)\frac{\partial \varphi_t}{\partial u_1} \wedge \frac{\partial \varphi_t}{\partial u_2}$ satisfies the differential equation

$$\dot{w}(t) = B(\varphi_t(u))w(t),$$

where

$$B = P_f P^{-1} + P\frac{\partial f}{\partial x}^{[2]} P^{-1},$$

where the matrix P_f is obtained by replacing each entry p_{ij} of P by its derivative in the direction of f, p_{ij_f}. The matrix $B = P_f P^{-1} + P\frac{\partial f}{\partial x}^{[2]} P^{-1}$ can be written in the following block form:

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

with

$$B_{11} = \begin{pmatrix} \begin{array}{c} \frac{\dot{B}_i}{B_i} - \frac{\dot{M}_i}{M_i} - \Big(\mu_M + \mu_B + \\ \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_i}{(B_i+B_u)^2}\Big) \end{array} & \frac{b_1\beta_2 M_i B_u}{(B_i+B_u)^2} \\ \frac{b_1\beta_2 M_i B_i}{(B_i+B_u)^2} & \begin{array}{c} \frac{\dot{B}_i}{B_i} - \frac{\dot{M}_i}{M_i} - \Big(\mu_M + \mu_B + \\ d_B + \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_2 M_i B_u}{(B_i+B_u)^2}\Big) \end{array} \end{pmatrix},$$

$$B_{12} = \begin{pmatrix} -\frac{B_i}{M_i}\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_u}{(B_i+B_u)^2} \\ -\frac{B_i}{M_i}\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i+B_u)^2} \end{pmatrix},$$

$$B_{21} = \begin{pmatrix} -\frac{M_i}{B_i}\frac{b_1\beta_2 B_u}{B_i+B_u} & -\frac{M_i}{B_i}\frac{b_1\beta_2 B_u}{B_i+B_u} \end{pmatrix},$$

$$B_{22} = -\left(2\mu_B + d_B + \frac{b_1\beta_2 M_i}{B_i + B_u}\right).$$

Let $z = (u, v, w)$ denote the vectors in $\mathbb{R}^3$; we select a norm in $\mathbb{R}^3$ as

$$|(u, v, w)| = \sup\{|u| + |v|, |w|\}.$$

Let $\mu(B)$ be the *Lozinskiĭ measure* of B with respect to the induced matrix norm $|\cdot|$ in $\mathbb{R}^3$, defined by

$$\mu(B) = \lim_{h\to 0^+} \frac{|I + hB| - 1}{h}.$$

Then the *Lozinskiĭ measure* $\mu(B)$ with respect to $|\cdot|$ can be estimated as follows (see [28] and [29]):

$$\mu(B) \le \sup\{\zeta_1, \zeta_2\},$$

where

$$\zeta_1 = \mu_1(B_{11}) + |B_{12}| \le \frac{\dot{B}_i}{B_i} - \frac{\dot{M}_i}{M_i} - \mu_M - \mu_B - \frac{b_1\beta_1 B_i}{B_i + B_u} + \frac{b_1\beta_1(\frac{\Pi_M}{\mu_M} - M_i)B_i}{(B_i + B_u)M_i}, \tag{3.18}$$

$$\zeta_2 = B_{22} + |B_{21}| \le \frac{b_1\beta_2 B_u M_i}{(B_i + B_u)B_i} - 2\mu_B - d_B - \frac{b_1\beta_2 M_i}{B_i + B_u}. \tag{3.19}$$

Note that $\mu_1(B_{11})$ is the *Lozinskiĭ measure* of the 2×2 matrix B_{11} with respect to the l_1 norm in $\mathbb{R}^2$, $|B_{12}|$ and $|B_{21}|$ are the operator norms of B_{12} and B_{21} when they are regarded as mappings from $\mathbb{R}$ to $\mathbb{R}^2$ and from $\mathbb{R}^2$ to $\mathbb{R}$, respectively, and $\mathbb{R}^2$ is endowed with the l_1 norm. Also note that since B_{22} is a scalar, its *Lozinskiĭ measure* with respect to any vector norm in $\mathbb{R}$ is equal to B_{22}.

Rewriting (3.1), we find that

$$\frac{b_1\beta_1(\frac{\Pi_M}{\mu_M}-M_i)B_i}{(B_i+B_u)M_i}=\frac{\dot{M}_i}{M_i}+\mu_M,\frac{b_1\beta_2B_uM_i}{(B_i+B_u)B_i}=\frac{\dot{B}_i}{B_i}+\mu_B+d_B. \tag{3.20}$$

From (3.18)–(3.20) we have

$$\mu(B)\le\frac{\dot{B}_i}{B_i}-\mu_B.$$

A solution $(M_i(t),B_u(t),B_i(t))$ to system (3.1) with $(M_i(0),B_u(0),B_i(0))$ in the absorbing set J exists for all $t>0$. Thus there exists $T>0$ such that $t>T$ implies that

$$\int_0^t\mu(B)dt\le\log\frac{B_i(t)}{B_i(0)}-\mu_Bt<-\frac{\mu_B}{2}t$$

for all $(M_i(0),B_u(0),B_i(0))\in J$. From a property of *Lozinskiĭ measure*, we have

$$\begin{aligned}S\varphi_t&=\int_{\overline{W}}\left|P(\varphi_t)\frac{\partial\varphi_t}{\partial u_1}\wedge\frac{\partial\varphi_t}{\partial u_2}\right|\\&\le\int_{\overline{W}}\left|P(\varphi)\frac{\partial\varphi}{\partial u_1}\wedge\frac{\partial\varphi}{\partial u_2}\right|\exp\left(\int_0^t\mu(B(\varphi_s(u))ds)\right)\\&\le S\varphi\exp\left(-\frac{\mu_B}{2}t\right).\end{aligned}$$

Therefore, $S\varphi_t\to0$ as $t\to\infty$. This contradicts (3.14), since ψ is invariant with respect to system (3.1), $\psi_t\in\Pi(\psi,\overline{U})$, and $\varphi_t(\overline{W})\subset J$ for all sufficiently large t. This contradiction implies that no simple closed rectifiable curve in $\overline{U}$ is invariant with respect to system (3.1). In particular, it rules out not only periodic trajectories but also homoclinic trajectories and heteroclinic loops, since each case gives rise to a simple closed rectifiable curve in $\overline{U}$.

We secondly claim that there is no period orbit on Σ. Suppose not, then there is a nontrivial periodic orbit χ in Σ, which must contain E^2 in its interior. Without loss of generality, we may assume that χ is the nearest periodic orbit from E^2 in Σ. This implies that $\chi\subset\overline{U}$, contradicting the above discussion. Similarly, we can conclude that there is no homoclinic orbit in Σ. Thus every positive limit set contains an equilibrium.

Since E_0 is locally asymptotically stable, let V be the basin of attraction on the equilibrium E_0. Pick up an initial point $Q=(M_i(0),B_u(0),B_i(0))$ and $Q\in\text{Int}\mathbb{R}^3_+\setminus(U\cup V)$. Then we can easily conclude that $\omega(Q)=\{E^1\}$, since every positive limit set contains an equilibrium, i.e., $Q\in W^s(E^1)$. This completes the proof of the first conclusion.

Second, let us consider the case $a_2>0$, $0<\frac{-a_1}{2a_2}<B^*_{i2}$, and $\Delta=0$. It follows from Theorem 2.1 that two positive equilibria E^1,E^2 coalesce into one positive equilibrium E^1. By Theorem 2.2, E_0 is locally asymptotically stable. This, together with the index theory of dynamics system on a two-dimensional compact manifold Σ which is positive invariant, implies that E^1 is unstable. Thus there is no period and homoclinic

orbit, and every positive limit set contains an equilibrium. Using the same way as in the proof of the first conclusion, we can conclude that the second conclusion is true.

Otherwise, by Theorem 2.1, there is no positive equilibrium on Σ. It follows from Proposition 3.3 that the disease-free equilibrium E_0 is globally asymptotically stable. This completes the proof of Theorem 3.8. $\square$

4. The classification for global behavior of system (2.2). In this section, we mainly study the dynamics of system (2.2). First, let us consider the dynamics of the following subsystem:

$$\tag{4.1} \begin{cases} \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_3 M_i S}{N_H} - \mu_H S, \\ \frac{dE}{dt} = \frac{b_2\beta_3 M_i S}{N_H} - \mu_H E - \alpha E, \\ \frac{dI}{dt} = \alpha E - \mu_H I - \delta I, \\ \frac{dH}{dt} = \delta I - \mu_H H - d_H H - \tau H, \\ \frac{dR}{dt} = \tau H - \mu_H R. \end{cases}$$

Combining Theorems 3.6, 3.7, and 3.8, all solutions of system (2.2) starting in $\mathcal{D}$ satisfy that $M_i(t) \to M_i^{\#}$ as $t \to +\infty$, where $M_i^{\#}$ is constant. Consequently, the limiting system of (4.1) is

$$\tag{4.2} \begin{cases} \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_3 M_i^{\#} S}{N_H} - \mu_H S, \\ \frac{dE}{dt} = \frac{b_2\beta_3 M_i^{\#} S}{N_H} - \mu_H E - \alpha E, \\ \frac{dI}{dt} = \alpha E - \mu_H I - \delta I, \\ \frac{dH}{dt} = \delta I - \mu_H H - d_H H - \tau H, \\ \frac{dR}{dt} = \tau H - \mu_H R. \end{cases}$$

After simple algebraic analysis, system (4.2) has a unique positive equilibrium $\tilde{E}^{\#}(S^{\#}, E^{\#}, I^{\#}, H^{\#}, R^{\#})$. The Jacobian matrix of the system (4.2) associated with $\tilde{E}^{\#}$ is the matrix A_{22} defined in section 2. Then, it follows from section 2 that the unique positive equilibrium $\tilde{E}^{\#}$ is locally asymptotically stable.

Using the change of variables $X_H = N_H + \frac{d_H}{\tau} R$, $X_1 = S + E$, and $X_2 = S + E + I$, system (4.2) can be written as

$$\tag{4.3} \begin{cases} \frac{dX_H}{dt} = \Pi_H - \mu_H X_H, \\ \frac{dX_1}{dt} = \Pi_H - (\mu_H + \alpha) X_1 + \alpha S, \\ \frac{dX_2}{dt} = \Pi_H - (\mu_H + \delta) X_2 + \delta X_1, \\ \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_2 M_i^{\#} S}{N_H} - \mu_H S, \\ \frac{dN_H}{dt} = \Pi_H - (\mu_H + d_H + \tau) N_H + d_H X_2 + \tau X_H. \end{cases}$$

It follows from the first equation of system (4.3) that $X_H(t) \to \frac{\Pi_H}{\mu_H}$ as $t \to +\infty$. Then, the limiting system of (4.3) is

$$(4.4)\qquad \begin{cases} \frac{dX_1}{dt} = \Pi_H - (\mu_H + \alpha)X_1 + \alpha S, \\ \frac{dS}{dt} = \Pi_H - \frac{b_2\beta_2 M_i^{\#} S}{N_H} - \mu_H S, \\ \frac{dN_H}{dt} = \Pi_H - (\mu_H + d_H + \tau)N_H + d_H X_2 + \tau\frac{\Pi_H}{\mu_H}, \\ \frac{dX_2}{dt} = \Pi_H - (\mu_H + \delta)X_2 + \delta X_1. \end{cases}$$

Straightforward calculation yields that system (4.4) has only one positive equilibrium $(X_1^{\#}, S^{\#}, N_H^{\#}, X_2^{\#}) = (S^{\#} + E^{\#}, S^{\#}, N^{\#}, S^{\#} + E^{\#} + I^{\#})$ in $\mathbb{R}_+^4$. From the Jacobian of system (4.4), we can easily verify that system (4.4) is a cooperative irreducible system in $\mathbb{R}_+^4$ [19]. By Theorem 3.1 in [19] or the result in [13], we conclude that the equilibrium $(X_1^{\#}, S^{\#}, N_H^{\#}, X_2^{\#})$ is globally asymptotically stable in $\mathbb{R}_+^4$. Since system (4.4) is the limiting system of (4.3) and system (4.3) is the limiting system of (4.2), it follows from Theorem 2.3 in paper [15] that the unique positive equilibrium $\tilde{E}^{\#}$ is a globally asymptotically stable equilibrium of system (4.2).

Therefore, the dynamics of system (2.2) can be easily derived from Theorems 3.6, 3.7, 3.8, and the above discussion.

THEOREM 4.1. (A) *If $R_0 > 1$ or $R_0 = 1$ and $\mu_M(d_B - \mu_B) > b_1\beta_1\mu_B$, then the unique positive equilibrium E^* of system* (2.2) *is globally asymptotically stable in $\mathcal{D}$.*

(B) *If $R_0 < 1$, then*

(B1) *if $a_2 > 0$, $0 < \frac{-a_1}{2a_2} < B_{i2}^*$, and $\Delta > 0$, then system* (2.2) *has switch phenomenon. That is, system* (2.2) *has a seven-dimensional stable manifold for E^1 which separates* $\text{int}\mathbb{R}_+^8$ *into two parts $V \times \mathbb{R}_+^5$ and $U \times \mathbb{R}_+^5$. If an initial point lies in $V \times \mathbb{R}_+^5$, then the solution $(M_i(t), B_u(t), B_i(t), S(t), E(t), I(t), H(t), R(t)) \to E_0$ as $t \to \infty$; if an initial point lies in $U \times \mathbb{R}_+^5$, then $(M_i(t), B_u(t), B_i(t), S(t), E(t), I(t), H(t), R(t)) \to E^2$ as $t \to \infty$; otherwise, the solution $(M_i(t), B_u(t), B_i(t), S(t), E(t), I(t), H(t), R(t)) \to E^1$ as $t \to \infty$;*

(B2) *if $a_2 > 0$, $0 < \frac{-a_1}{2a_2} < B_{i2}^*$, and $\Delta = 0$, then system* (2.2) *also has a seven-dimensional stable manifold for E^1 which also separates* $\text{int}\mathbb{R}_+^8$ *into two parts. If an initial point lies in $V \times \mathbb{R}_+^5$, then the solution $(M_i(t), B_u(t), B_i(t), S(t), E(t), I(t), H(t), R(t)) \to E_0$ as $t \to \infty$; otherwise, the solution $(M_i(t), B_u(t), B_i(t), S(t), E(t), I(t), H(t), R(t)) \to E^1$ as $t \to \infty$.*

(C) *Otherwise, the disease-free equilibrium E_0 is globally asymptotically stable.*

5. Discussion. In this section, we mainly summarize our results and make some further remarks.

In this article, we have mainly studied the asymptotical behavior of a nine-dimensional WNV model. By applying an asymptotically autonomous convergence theorem, competitive and cooperative theory, and a criterion for the orbital stability of periodic orbits associated with higher-dimensional nonlinear autonomous systems, we provide a complete classification of dynamics for this model in Theorem 4.1, which includes the critical case. It is shown that the dynamics of system (2.2) is determined by the basic reproduction number and the number of the positive equilibria. There are only four cases:

(1) If there is no positive equilibrium, then the unique disease-free equilibrium E_0 is globally asymptotically stable. In this case, the dynamics of the model on Σ can be depicted in Figure 1(a).

(2) If $R_0 \geq 1$ and system (2.2) has a unique positive equilibrium, then the unique endemic equilibrium is globally asymptotically stable in the interior of the feasible

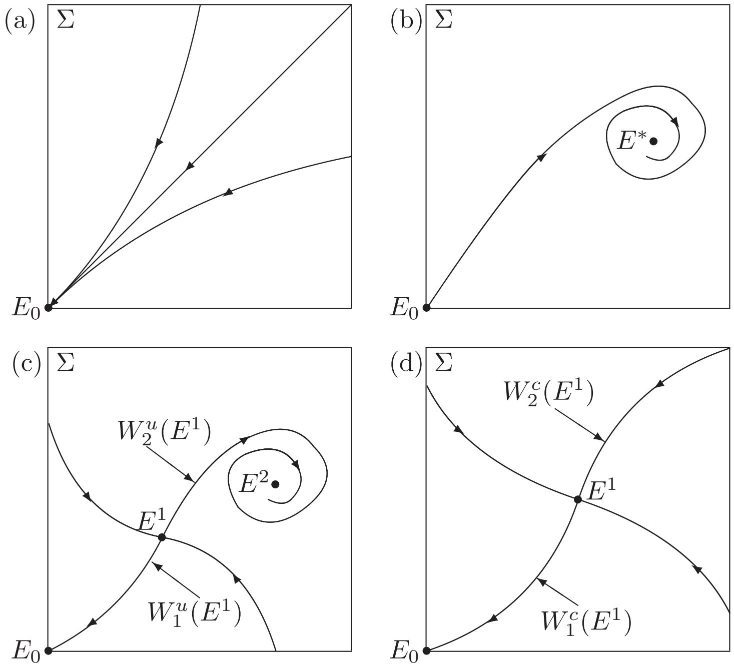


FIG. 1. *The possible dynamics of system* (2.2) *on* Σ.

region, and the disease persists at an endemic equilibrium if it initially exists. In particular, when $R_0 > 1$, system (2.2) has a unique positive equilibrium, and the unique positive is globally asymptotically stable in $\mathrm{Int}\mathbb{R}^8_+$. This result provides a positive answer for the conjecture proposed in [11, 12]. In this case, the dynamics of the model on Σ can be depicted in Figure 1(b).

(3) If system (2.2) has two positive equilibria E^1, E^2, then system (2.2) has switch phenomenon, that is, a seven-dimensional stable manifold for E^1 separates $\mathrm{int}\mathbb{R}^8_+$ into two parts. One contains E_0 and one branch $W_1^u(E^1)$ of the unstable manifold for E^1, the other contains E^2 and the other branch $W_2^u(E^1)$ of the unstable manifold for E^1. All solutions not in $W^s(E^1)$ are convergent to either E_0 or E^2. This result indicates that the reproduction number can't simply describe whether WNV will prevail or not and suggests that we should pay attention to the initial states of WNV. In this case, the dynamics of the model on Σ can be depicted in Figure 1(c).

(4) Otherwise, system (2.2) has a unique positive equilibrium E^1 which has one-dimensional center manifold, and there exists a seven-dimensional stable manifold for E^1 which also separates $\mathrm{int}\mathbb{R}^8_+$ into two parts. One contains E_0 and one branch $W_1^c(E^1)$ of the center manifold for E^1, the other contains no positive equilibrium and the other branch $W_2^c(E^1)$ of the center manifold for E^1. All solutions are convergent to either E_0 or E^1, respectively. In this case, the dynamics of system (2.2) on Σ can be depicted in Figure 1(d).

The basic reproduction number R_0 of the model can be expressed by

$$R_0 = \sqrt{\frac{b_1^2\beta_1\beta_2\mu_B\Pi_M}{\mu_M^2(\mu_B + d_B)\Pi_B}}.$$

Now let us explain the biological means of R_0. If system (1.1) has no infected female mosquitoes and infected birds and is in balance, then Theorem 2.1 implies that the numbers of the susceptible female mosquitos and birds are $\frac{\Pi_M}{\mu_M}$ and $\frac{\Pi_B}{\mu_B}$, respectively. On these conditions, the average times that one infected bird is bitten by mosquitoes can be defined by

$$\hat{b}_1 := \frac{b_1 \frac{\Pi_M}{\mu_M}}{\frac{\Pi_B}{\mu_B}}.$$

Since the mean lifespan of the infected bird is

$$\hat{\tau}_B := \frac{1}{\mu_B + d_B},$$

the average number of the infected female mosquitoes generated by infection of the infected bird can be defined as

$$\hat{M}_i := \hat{b}_1 \beta_1 \hat{\tau}_B = \frac{b_1 \beta_1 \frac{\Pi_M}{\mu_M}}{\frac{\Pi_B}{\mu_B}} \frac{1}{\mu_B + d_B}.$$

The average number of the infected birds generated by infection of the infected female mosquitoes can be defined as

$$\hat{B}_i := b_2 \beta_2 \frac{1}{\mu_M}.$$

Therefore, the total number of secondary cases generated by transmission of one infected bird can be defined by

$$\tilde{R_0} := \hat{B}_i \hat{M}_i = \frac{b_1 \beta_1 \frac{\Pi_M}{\mu_M}}{\frac{\Pi_B}{\mu_B}} \frac{1}{\mu_B + d_B} b_2 \beta_2 \frac{1}{\mu_M} = (R_0)^2$$

when the numbers of the susceptible mosquitoes and birds are $\frac{\Pi_M}{\mu_M}$ and $\frac{\Pi_B}{\mu_B}$, respectively. If $R_0 > 1$, i.e., $\tilde{R_0} > 1$, Theorem 4.1(A) implies that the WNV will prevail, since an infective bird will be replaced with greater than one new case. If $R_0 < 1$ is less than one, Theorem 4.1(B) implies that the WNV will be likely to fade out, since an infective individual will be replaced with less than one new case.

It follows from the expression of R_0 that the quantity R_0 grows with the recruitment of uninfected mosquito population, the natural death rate of birds, and the per capita rate of the mosquitoes on the birds but falls with the natural death rate of mosquitoes, the recruitment of uninfected birds, and the WNV-induced death rate of birds. Thus, this result implies that the WNV spreads more rapidly if birds migrate to a region with higher mosquito density, and it also implies that it is an efficient way to halt the spread of WNV by using the mosquito-reduction strategies. However, the decreasing of the recruitment of the uninfected birds and the increasing natural death rate of the birds is beneficial to the prevalence of the WNV. This tells us that it is a risk factor for the spread of WNV to kill the birds during the period that the WNV prevails. By contraries, we should increase the recruitment of the uninfected birds.

Additionally, in this paper we consider only the dynamics of the model with WNV transmission among one single mosquito–bird-human community. However, the effect of seasonality and migration of birds are important factors, since, for nontropical regions, the cold season signals the end of the epidemic season, while the heterogeneity and migration of birds from region to region plays a key role in the viral amplification process. Also, it is feasible that infected birds can migrate from one region to another. Thus, it is also interesting to study how the dynamics of the model incorporates infected birds can migrate from one region to another. We leave these for future investigations.

Acknowledgments. The problem for global stability of the unique equilibrium was first proposed to the first author by Prof. Wu Jianhong in the summer of the year 2004 when he visited Prof. Wu at York University. The first author greatly appreciates Prof. Wu for his encouragement and many good suggestions during studying the problem for past years. Both authors are also grateful to Prof. Zhu Huaiping for his biological explanation for this model.

REFERENCES

[1] J. Arino, C. C. Mccluskey, and P. Van den Driessche, *Global results for an epidemic model with vaccination that exhibits backward bifurcation,* SIAM J. Appl. Math., 64 (2003), pp. 260–276.

[2] W. D. Wang and Z. Ma, *Global dynamics of an epidemic model with time delay,* Nonlinear Anal. Real World Application, 3 (2002), pp. 365–373.

[3] W. D. Wang and X. Q. Zhao, *An epidemic model in a patchy environment,* Math. Biosci., 190 (2004), pp. 97–112.

[4] Centers for Disease Control and Prevention, *West Nile Virus: Fact Sheet*, http://www.cdc.gov/ncidod/dvbid/westnile/wnv factSheet.htm.

[5] Center for Disease Control and Prevention, *CDC-West Nile Virus-surveillance and Control Case Count of West Nile Disease,* http://www.cdc.gov/ncidod/dvbid/westnile/.

[6] C. C. Lord and J. F. Day, *Simulation studies of St. Louis encephalitis and West Nile viruses: The impact of bird mortality,* Vector Borne and Zoonotic Diseases, 1 (2001), pp. 317–329.

[7] D. M. Thomas and B. Urena, *A model describing the evolution of West Nile-like encephalitis in New York City,* Math. Comput. Modelling, 34 (2001), pp. 771–781.

[8] M. J. Wonham, T. de-Camino-Beck, and M. Lewis, *An epidemiological model for West Nile virus: Invasion analysis and control applications,* Proc. R. Soc. Lond. Ser. B, 1538 (2004), pp. 501–507.

[9] V. M. Kenkre, R. R. Parmenter, I. D. Peixoto, and L. Sadasiv, *A theoretical framework for the analysis of the West Nile virus epidemic,* Math. Comput. Modelling, 42 (2005), pp. 313–324.

[10] R. S. Liu, J. P. Shuai, J. Wu, and H. Zhu, *Modeling spatial spread of West Nile virus and impact of directional dispersal of birds,* Math. Biosci. Eng., 3 (2006), pp. 145–160.

[11] C. Bowman, A. B. Gumel, J. Wu, P. Van den Driessche, and H. Zhu, *A mathematical model for assessing control strategies against West Nile virus,* Bull. Math. Biol., 67 (2005), pp. 1107–1133.

[12] J. F. Jiang, Z. P. Qiu, J. Wu, and H. Zhu, *Threshold conditions for West Nile virus outbreaks,* Bull. Math. Biol., 2008, DOI 10.1007/s11538-008-9374-6.

[13] J. F. Jiang, *On the global stability of cooperative systems,* Bull. London Math. Soc., 6 (1994), pp. 455–458.

[14] C. Castillo-Chavez, W. Huang, and J. Li, *Competitive exclusion in gonorrhea models and other sexually transmitted diseases,* SIAM J. Appl. Math., 56 (1996), pp. 494–508.

[15] C. Castillo-Chavez and H. R. Thieme, *Asymptotically autonomous epidemic models,* in Mathematical Population Dynamics: Analysis of Heterogeneity, Vol. 1, O. Arino, D. Axelrod, M. Kimmel, M. Langlais, eds., Wuerz, Winnipeg, 1995, pp. 33–50.

[16] H. R. Thieme, *Convergence results and a Poincaré-Bendixson trichotomy for asymptotically autonomous differential equations,* J. Math. Biol., 30 (1992), pp. 755–763.

[17] H. R. Thieme, *Asymptotically autonomous differential equations in the plane,* Rocky Mountain J. Math., 24 (1994), pp. 351–380.

[18] P. Van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission,* Math. Biosci., 180 (2002), pp. 29–48.

[19] H. L. Smith, *Monotone Dynamical Systems: An Introduction to Theory of Competitive and Cooperative Systems,* Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[20] Y. Wang and J. F. Jiang, *The general properties of discrete time competitive dynamical systems,* J. Differential Equations, 176 (2001), pp. 470–493.

[21] P. Takac, *Domains of attraction of generic ω-limit sets for strongly monotone discrete-time semigroups,* J. Reine Angew. Math., 423 (1992), pp. 101–173.

[22] D. London, *On derivations arising in differential equations,* Linear Multilinear Algebra, 4 (1976), pp. 179–189.

[23] J. S. Muldowney, *Compound matrices and ordinary differential equation,* Rocky Mountain J. Math., 20 (1990), pp. 857–872.

[24] M. Y. Li and J. S. Muldowney, *Global stability for the SEIR model in epidemiology,* Math. Biosci., 125 (1995), pp. 155–164.

[25] H. R. Thieme, *Persistence under relaxed point-dissipativity (with an application to an endemic model),* SIAM J. Math. Anal., 24 (1993), pp. 407–435.

[26] M. Y. Li and J. Muldowney, *On Bendixson's crition,* J. Differential Equations, 106 (1994), pp. 27–39.

[27] M. Y. Li and J. Muldowney, *A geometric approach to global-stability problems,* SIAM J. Math. Anal., 27 (1996), pp. 1070–1083.

[28] R. H. Martin, Jr., *Logarithmic norms and projections applied to linear differential systems,* J. Math. Anal. Appl., 45 (1974), pp. 432–454.

[29] J. S. Muldowney, *Dichotomies and asymptotic behavior for linear differential systems*, Trans. Amer. Math. Soc., 283 (1984), pp. 465–484.

[30] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semiflows: Chain recurrence and Lyapunov functions,* Trans. Amer. Math. Soc., 347 (1995), pp. 1669–1685.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1228–1243

STABILITY OF CURVED INTERFACES IN THE PERTURBED TWO-DIMENSIONAL ALLEN–CAHN SYSTEM*

DAVID IRON[†], THEODORE KOLOKOLONIKOV[†], JOHN RUMSEY[†], AND JUNCHENG WEI[‡]

Abstract. We consider the singular limit of a perturbed Allen–Cahn model on a bounded two-dimensional domain: $\begin{cases} u_t = \varepsilon^2 \Delta u - 2(u-\varepsilon a)(u^2-1), & x \in \Omega \subset \mathbb{R}^2 \\ \partial_n u = 0, & x \in \partial\Omega \end{cases}$ where ε is a small parameter and a is an $O(1)$ quantity. We study equilibrium solutions that have the form of a curved interface. Using singular perturbation techniques, we fully characterize the stability of such an equilibrium in terms of a certain geometric eigenvalue problem, and give a simple geometric interpretation of our stability results. Full numerical computations of the time-dependent PDE as well as of the associated two-dimensional eigenvalue problem are shown to be in excellent agreement with the analytical predictions.

Key words. Allen–Cahn equation, interface motion, spectral analysis, matched asymptotic expansions

AMS subject classifications. 35B25, 35B32, 35K57

DOI. 10.1137/070706380

1. Introduction.

We consider a perturbed two-dimensional Allen–Cahn equation,

$$\begin{cases} u_t = \varepsilon^2 \Delta u + f(u) + \varepsilon g(u), & x \in \Omega \subset \mathbb{R}^2\,, \\ \partial_n u = 0, & x \in \partial\Omega\,. \end{cases} \tag{1}$$

Here, Ω is a smooth two-dimensional domain and $f(u)$ is a smooth function having the following properties:

1. f has three roots $u_- < u_0 < u_+$ with $f'(u_\pm) < 0$,
2. $\int_{u_-}^{u_+} f(u)\,du = 0$,

and $g(u)$ is any smooth function function with $\int_{u_-}^{u_+} g(u)\,du \neq 0$.

The standard Allen–Cahn equation corresponds to $g = 0$, $f = -2u(u^2-1)$. This model was introduced in [2] as a simple model of evolution of antiphase boundaries and is now well understood. In the limit $\varepsilon \to 0$, the solution forms a sharp interface layer. On one side of the interface, $u \sim u_-$, while on the other, $u \sim u_+$. Once the interface layer is formed, its motion is described by the mean curvature law which minimizes the perimeter of the interface ([5], [9]). The stable stationary solution corresponds to an interface with a minimal perimeter that intersects the boundary orthogonally ([12]). Therefore, any nontrivial stable steady equilibrium of the unperturbed Allen–Cahn equation consists of a straight interface. The stability of such an interface has been analyzed by several authors in variety of settings; see for instance [1], [10], [11], [14], [16], [18]. The main result is that such an interface can be stable provided the domain contains a "neck". More precisely, as shown in [10], [11], in the limit $\varepsilon \to 0$,

*Received by the editors October 25, 2007; accepted for publication (in revised form) September 26, 2008; published electronically February 4, 2009.
http://www.siam.org/journals/siap/69-5/70638.html

[†]Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada (iron@mathstat.dal.ca, tkolokol@mathstat.dal.ca, John.Rumsey@mathstat.dal.ca).

[‡]Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk).

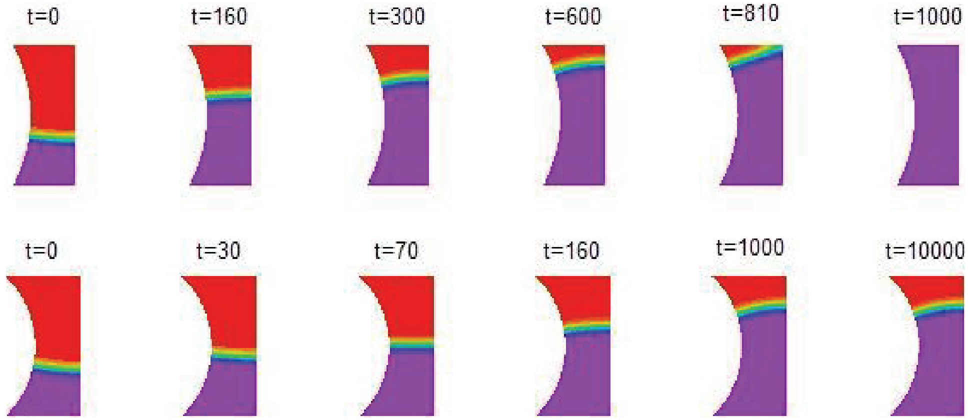


FIG. 1. *Motion of an interface for the perturbed Allen–Cahn model given by* $u_t = \varepsilon^2 \Delta u - 2(u - \varepsilon a)(u^2 - 1)$*, with* $a = 0.3, \varepsilon = 0.07$*. Top row: the interface is unstable and eventually disappears. Bottom row: The interface gets "stuck" in the middle of the domain; a nontrivial equilibrium is reached. The domain height is* 1.5 *and the distance between the side boundaries is* 0.5*. The radius of the left boundary is* 1.5 *for the top row and* 1.0 *for the bottom row.*

the interface stability depends only on the curvatures κ_+, κ_- of the boundary at the two points that intersect the interface, and the interface length ℓ. The interface is stable provided that $\ell + \kappa_+^{-1} + \kappa_-^{-1} < 0$. Geometrically, the threshold case corresponds to the two boundaries that are locally concentric.

More generally, the perturbed Allen–Cahn equation (1) is used as a prototype model of wave propagation in various contexts. In two or higher dimensions, a small perturbation leads to *weakly curved fronts*. For an overview, see [15], Chapter 2.2. In the absence of boundaries, the front becomes a closed curve which lies on a perimeter of some circle. Moreover, such a front is unstable, and either shrinks to a point or else expands indefinitely, depending on the initial conditions [15]. A typical nonlinearity is $f + \varepsilon g = -2(u - A)(u^2 - 1)$ where A is close to 0. This system (but without the assumption that A is small) was used as a simple model of spreading depressions in the human brain that are associated with cerebral strokes [4]. (When A is replaced by an inhomogeneous term $a(x)$, it is called the Fife–Greenlee problem [8], [6].) For convex domains, it is known ([3] [13]) that the only stable solution is a trivial equilibrium. Indeed any interface propagates until it merges with the boundary and disappears.

However, when the domain consists of two boxes of different heights, it was shown in [4] that the interface can get "stuck" at the juncture between the two boxes, provided their dimensions are sufficiently different. A similar phenomenon was reported in [17], where the propagation of chemical pulses in complex geometries with corners and junctures was studied numerically and experimentally.

The perturbation by a small term $\varepsilon g(u)$ has a large effect on the shape and stability of the interface. In particular, the equilibrium solution now consists of a *curved* interface. In the limit $\varepsilon \to 0$, this curve is part of a circular arc whose radius $\hat{R}$, given by (2) below and is independent of the domain shape. For non-convex domains, it is possible to get a stable interface. One such domain is illustrated in Figure 1. It consists of a rectangle with a circular cutout. In the first simulation (top row), the interface propagates through the domain without reaching any equilibrium, whereas in the second simulation (second row) the interface settles to a steady state somewhere in the middle of the domain. The only difference between the two simulations is the curvature of the left boundary of the domain, which has been increased in the second simulation.

In this paper, we fully characterize the stability of curved interfaces. First, we provide the necessary and sufficient conditions that describe the stability of an interface. Second, we give a simple geometric interpretation of our stability results.

Before stating our stability result, we characterize the radius of the steady state. This simple result was already given in [16], Appendix A. We summarize it here as following.

Proposition 1. *Let U be a solution to*

$$U''(y) + f(U) = 0, \qquad U \to u_{\pm} \quad \text{as } y \to \pm\infty$$

and define

$$\hat{R} = -\frac{\int_{-\infty}^{\infty} U'^2(y)dy}{\int_{u_-}^{u_+} g(u)du}. \tag{2}$$

Suppose that there exists a circle of radius $\hat{R}$ which intersects $\partial\Omega$ orthogonally, and let p be its center. Then in the limit $\varepsilon \to 0$ we have

$$u(x) \sim U\left(\frac{\hat{R} - |p - x|}{\varepsilon}\right), \quad \varepsilon \to 0. \tag{3}$$

Moreover, any solution to (1) *of the form* (3) *must satisfy* (2).

We are now ready to state our main result.

Theorem 2. *Let $u(x)$ be the steady-state solution as given in Proposition* 1 *and $\hat{R}$ its radius as defined in* (2). *Let ℓ be the length of the interface and let κ_+, κ_- be the curvatures of the boundary at the points which intersect the interface. Consider the stability problem associated with* (1),

$$\begin{cases} \lambda\phi = \varepsilon^2\Delta\phi + f'(u)\phi + \varepsilon g'(u)\phi, & x \in \Omega \\ \partial_n\phi = 0, \quad x \in \partial\Omega. \end{cases} \tag{4}$$

In the limit $\varepsilon \to 0$, the eigenvalues λ are of $O(\varepsilon^2)$ given by

$$\lambda = \varepsilon^2\lambda_0, \tag{5a}$$

where λ_0 solves the following geometric eigenvalue problem*:*

$$\begin{cases} T'' + \left(\hat{R}^{-2} - \lambda_0\right)T = 0 \\ T'(-\ell/2) + \kappa_- T(-\ell/2) = 0 \\ T'(\ell/2) - \kappa_+ T(\ell/2) = 0. \end{cases} \tag{5b}$$

Thus, the interface is stable if all solutions λ_0 of (5b) *are negative, and unstable if at least one solution is positive. Equivalently, λ_0 solves*

$$\lambda_0 = \frac{1}{\hat{R}^2} - \mu^2 \quad \text{where} \quad \tan(\mu\ell) = -\frac{\mu(\kappa_+ + \kappa_-)}{\mu^2 - \kappa_+\kappa_-} \tag{6}$$

or

$$\arctan\left(\frac{-\kappa_+}{\mu}\right) + \arctan\left(\frac{-\kappa_-}{\mu}\right) = \mu\ell \tag{7}$$

for some branch of arctan.

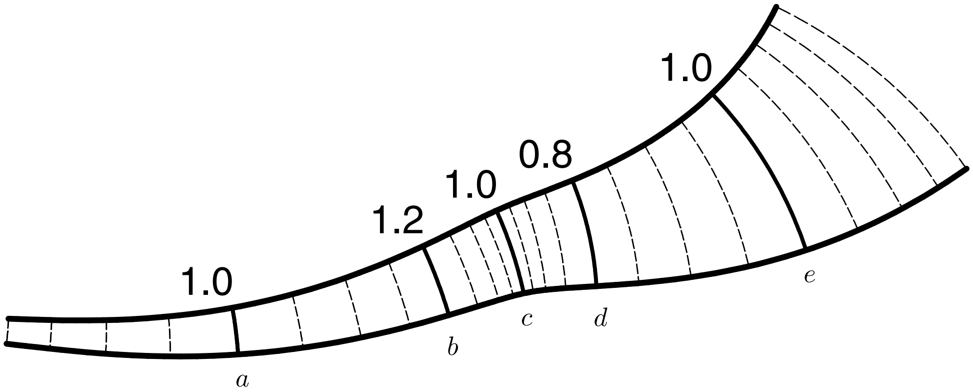


Fig. 2. *Geometric interpretation of stability criterion (see Theorem* 3*). The numbers indicate the radius of the corresponding interface below that number. The maximum and minimum radius is* 1.2 *and* 0.8*, respectively. If* $\hat{R} = 1$*, then curve c represents the location of a stable interface, whereas curves a and e correspond to unstable interfaces.*

Remark. Suppose that $\lambda_0 \neq 0$, i.e., the geometric eigenvalue problem (5b) has no zero eigenvalue. Then the existence of such steady state can be rigorously proved, following the lines of [11]. We omit the details.

In the case of the unperturbed Allen–Cahn equation ($g = 0, \quad \hat{R} = \infty$), the geometric eigenvalue problem (5b) is identical to (1.5) obtained by Kowalczyk in [10], [11]. However, here we use a somewhat different method using solvability condition and test functions.

The stability criterion (5b) has a natural geometric interpretation which we now discuss. Consider a domain such as shown in Figure 2. Parameterize the top boundary in terms of arclength s, from left to right, and let $q(s)$ be the corresponding point on the top boundary. We suppose that there is a unique circle that goes through $q(s)$ and that intersects both top and bottom boundaries orthogonally. Let $R(s)$ denote the radius of such a circle. Then we have the following.

Theorem 3. *Let $\hat{R}$ be the radius of a steady interface as defined in Proposition* 1*, let $R(s)$ be as defined above, and suppose that $R(s) = \hat{R}$ for some s. Then the interface is stable if $R'(s) < 0$ and it is unstable if $R'(s) > 0$.*

For example, for the domain as shown in Figure 2, if $\hat{R} \in (0.8, 1.2)$, then there exists a stable steady interface between curves b and d. On the other hand, any interface to the left of b or to the right of d is unstable. To our knowledge, this is the first result that combines both the effects of perturbation and the effects of the boundary.

The rest of the paper is outlined as follows. Proposition 1 is derived in section 2. The main result, Theorem 2, is then derived in section 3. Finally we prove Theorem 3 in section 4. We conclude with numerical calculations in section 5 and some discussions and open problems in section 6.

2. Equilibrium front solution.

In this section we construct the steady state consisting of a single interface. The main goal is to derive (2) of Proposition 1.

We seek a solution which divides the domain into two regions. In one of the regions $u \sim u_+$ and in the other $u \sim u_-$. The two regions are separated by an interface, or front, of thickness $O(\varepsilon)$. We expect the interface to be localized about a circle segment which intersects the boundary of Ω orthogonally. Let $\hat{R}$ be the radius

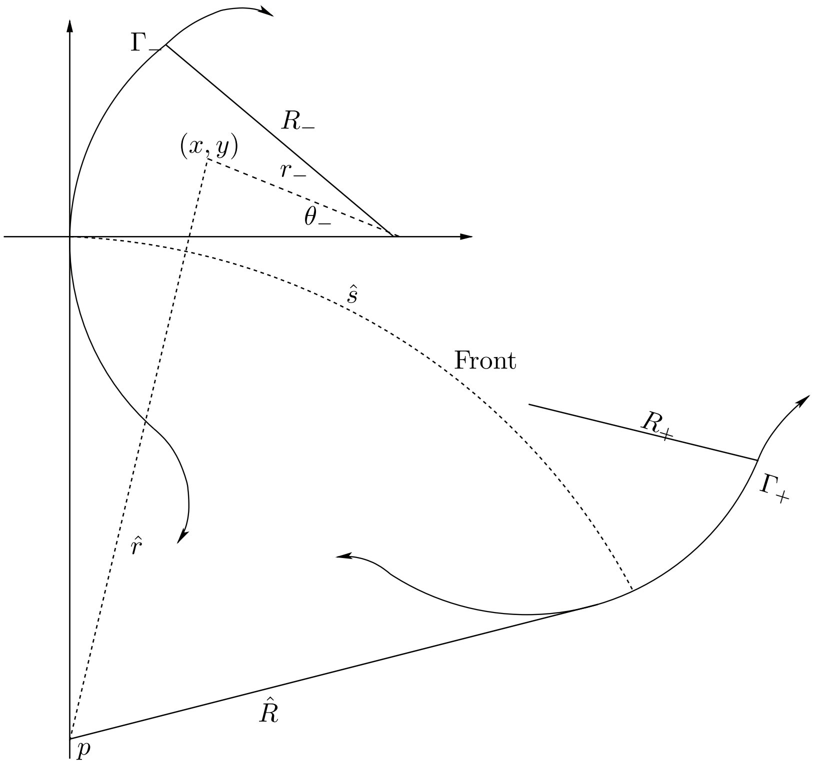


FIG. 3. *Schematic used for the derivation of coordinate systems in the interior of the domain and localized near the boundaries.*

of the interface and define the following coordinate system as illustrated in Figure 3:

$$x = R_- - r_- \cos(\theta_-) = \hat{r}\sin(\hat{s}/\hat{R})\,, \tag{8}$$

$$y = r_- \sin(\theta_-) = \hat{r}\cos(\hat{s}/\hat{R}) - \hat{R}\,. \tag{9}$$

Near the boundaries, we define localized coordinates $\rho_\pm$ and $t_\pm$ as follows:

$$\rho_\pm \equiv \frac{r_\pm - R_\pm}{\varepsilon}\,, \qquad t_\pm \equiv \frac{R_\pm \theta_\pm}{\varepsilon}\,. \tag{10}$$

Here, + and − are used to denote the right and left curved boundaries, respectively. The ± will be dropped whenever the meaning is clear. We also define coordinates localized near the front by

$$\hat{\rho} \equiv \frac{\hat{r} - \hat{R}}{\varepsilon}\,. \tag{11}$$

We can then write $\hat{\rho}$ as a function of t and ρ:

$$\hat{\rho} = t - \varepsilon\left(\frac{\rho t}{R} - \frac{\rho^2}{2\hat{R}}\right) + \cdots\,. \tag{12}$$

In the interior of the domain, we expect the front to be radially symmetric. Thus, in the new coordinate system, the equilibrium front will satisfy

$$u_{\hat{\rho}\hat{\rho}} + \frac{\varepsilon}{\hat{R} + \varepsilon\hat{\rho}} u_{\hat{\rho}} + f(u) + \varepsilon g(u) = 0\,, \tag{13}$$

in the interior of the domain. We expand

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \cdots\,, \tag{14}$$

substitute into (13), and collect powers of ε to obtain,

$$u_0'' + f(u_0) = 0\,, \tag{15}$$

$$u_1'' + f'(u_0)u_1 = -\frac{1}{\hat{R}}u_0' - g(u_0)\,, \tag{16}$$

$$u_2'' + f'(u_0)u_2 = \frac{1}{\hat{R}^2}\hat{\rho}u_0' - \frac{1}{\hat{R}}u_1' - \frac{f''(u_0)u_1^2}{2} - g'(u_0)u_1\,. \tag{17}$$

From here on $'$ denotes differentiation with respect to $\hat{\rho}$ when associated with u_i. In all other cases $'$ will represent differentiation with respect to the appropriate argument. At this point it is convenient to define the operator $L\psi \equiv \psi'' + f'(u_0)\psi$.

From conditions 1 and 2 following (1), u_0 will be given by the unique heteroclinic orbit connecting u_+ to u_-. For the case $f(u) = 2u(1-u^2)$, we have the exact solution $u_0 = \tanh(\hat{\rho})$. We note that by differentiating (15) with respect to $\hat{\rho}$, $Lu_0' = 0$.

To determine $\hat{R}$, we consider the steady-state system,

$$\varepsilon^2 \Delta u + f(u) + \varepsilon g(u) = 0\,. \tag{18}$$

We multiply (18) by u_0' and integrate over the domain,

$$\int_\Omega u_0'(\varepsilon^2 \Delta u + f(u) + \varepsilon g(u))\, dA = 0\,. \tag{19}$$

Applying Green's identity to (19) we obtain

$$-\varepsilon^2 \int_{\partial\Omega} u\, \partial_n u_0'\, ds + \int_\Omega \varepsilon^2 u\, \Delta(u_0') + u_0'(f(u) + \varepsilon g(u))\, dA = 0\,. \tag{20}$$

We now use (14) and (11) in (20) and collect powers of ε to obtain

$$\begin{aligned} -\varepsilon^2 \int_{\partial\Omega} u_0 \partial_n u_0'\, ds + \int_\Omega \Bigg(&(u_0(u_0')'' + f(u_0)u_0') + \varepsilon \\ &\left((u_0')''u_1 + \frac{1}{\hat{R}}(u_0')'u_0 + f'(u_0)u_1u_0' + g(u_0)u_0' \right) \Bigg)\, dA = 0\,. \end{aligned} \tag{21}$$

Integrating over $\hat{\rho}$ by parts and using $\lim_{\hat{\rho}\to\pm\infty} u_0' = 0$ yields

$$\int_\Omega (u_0')''u_0\, dA = \int_\Omega u_0''u_0'\, dA\,, \tag{22}$$

$$\int_\Omega (u_0')'u_0\, dA = \int_\Omega (u_0')^2\, dA\,. \tag{23}$$

Using (15) and $Lu_0 = 0$, (21) may be written as

$$-\varepsilon \int_{\partial\Omega} \partial_n u_0' u_0\, ds = -\int_\Omega \left(\frac{u_0'}{\hat{R}} + g(u_0) \right) u_0'\, dA\,. \tag{24}$$

Using (12) we find the leading order behavior of $\partial_n u_0'|_{\partial\Omega}$:

$$\partial_n u_0'|_{\partial\Omega} \sim \frac{\partial}{\partial\rho} u_0' \left(t - \varepsilon \left(\frac{\rho t}{R} - \frac{\rho^2}{2\hat{R}} \right) \right) \bigg|_{\rho=0} , \tag{25}$$

$$= -\varepsilon u_0'' \frac{t}{R} . \tag{26}$$

Thus, the boundary term in (24) is of a much lower order, and the equilibrium radius of the front is given by

$$\hat{R} \sim - \frac{\int_{-\infty}^{\infty} (u_0')^2 \, dt}{\int_{u_-}^{u_+} g(y) \, dy} . \tag{27}$$

This shows, that to leading order, $\hat{R}$ is independent of the domain shape and completes the derivation of Proposition 1.

3. Proof of Theorem 2. We now construct a solvability condition to determine the principal eigenvalues of (4). Since u_0' is of one sign and $Lu_0' = 0$, we expect that the principal eigenvalue is small and to leading order the principal eigenfunction will behave like u_0' in the interior of the domain. Such an eigenfunction is often referred to as a translation eigenfunction as it is associated with the near translation invariance of the front in the interior of the domain with respect to the radial co-ordinate. In this case, $\hat{s}u_0'$ also satisfies (4) to leading order and as a result, we will need two solvability conditions to determine the principal eigenvalue.

We construct our solvability conditions by multiplying (4) by test function v and integrating over the domain we obtain

$$\int_\Omega v \left(\varepsilon^2 \Delta\phi + f'(u)\, \phi \right) dA + \varepsilon \int g'(u)\, \phi \, v \, dA = \lambda \int_\Omega \phi \, v \, dA , \tag{28}$$

where v is of the form

$$v(\hat{s}, \hat{\rho}) = w(\hat{s}) \, u_0'(\hat{\rho}) \tag{29}$$

and $w(\hat{s})$ is an arbitrary test function.

Using Green's identity and applying the boundary conditions in (4) results in

$$-\varepsilon^2 \int_{\partial\Omega} \phi \, \partial_n v \, ds + \int_\Omega \Big(\varepsilon^2 \Delta v + f'(u)\, v + \varepsilon \, g'(u)\, v \Big) \phi \, dA = \lambda \int_\Omega \phi \, v \, dA . \tag{30}$$

Here, s is arc length along the boundary and dA is an element of area in the interior. From (10) and (11),

$$ds = R \, d\theta = \varepsilon \, dt , \tag{31}$$

$$dA = \frac{\hat{r}}{\hat{R}} \, d\hat{r} \, d\hat{s} = \varepsilon \left(1 + \varepsilon \frac{\hat{\rho}}{\hat{R}} \right) d\hat{\rho} \, d\hat{s} . \tag{32}$$

Consider the $\int_\Omega \Big(\varepsilon^2 \Delta v + f'(u)\, v + \varepsilon \, g'(u)\, v \Big) \phi \, dA$ term in (30), in which v, u, ϕ are written in the interior coordinates $\hat{r}$ and $\hat{s}$. Expand ϕ:

$$\phi = \phi_0 + \varepsilon \phi_1 + \varepsilon^2 \phi_2 + \cdots . \tag{33}$$

Use (11), (32), (33), and (14) to write $\int_\Omega \Big(\varepsilon^2\Delta v + f'(u)\,v + \varepsilon\, g'(u)\,v\Big)\phi\,dA$ in terms of the coordinates, $\hat\rho$ and $\hat s$:

$$\begin{aligned}
&\Big(\varepsilon^2\Delta v(\hat r) + f'(u)\,v(\hat r) + \varepsilon\, g'(u)\,v(\hat r)\Big)\phi\,dA\\
&\quad\sim \left[\varepsilon^2\left(\frac{1}{\varepsilon^2}\,v_{\hat\rho\hat\rho} + \frac{1}{\hat R+\varepsilon\hat\rho}\,\frac{1}{\varepsilon}v_{\hat\rho} + v_{\hat s\hat s}\right) + f'\left(u_0+\varepsilon\,u_1+\varepsilon^2\,u_2\right)\,v\right.\\
&\qquad\quad \left. +\ \varepsilon\, g'\left(u_0+\varepsilon\,u_1+\varepsilon^2\,u_2\right)\,v\right]\Big[\phi_0+\varepsilon\phi_1+\varepsilon^2\phi_2\Big]\left[\varepsilon\left(1+\varepsilon\,\frac{\hat\rho}{\hat R}\right)d\hat\rho\,d\hat s\right]\\
&\quad\sim \left\{\varepsilon^2\left[\frac{1}{\hat R}\,v_{\hat\rho}\,\phi_0 + u_1\,f''(u_0)\,v\,\phi_0 + g'(u_0)\,v\,\phi_0\right]\right.\\
&\qquad\quad + \varepsilon^3\Bigg[v_{\hat s\hat s}\phi_0 + u_2\,f''(u_0)\,v\,\phi_0 + \frac{1}{2}\,u_1^2\,f'''(u_0)\,v\,\phi_0 + u_1\,g''(u_0)\,v\,\phi_0\\
&\qquad\qquad + \frac{1}{\hat R}\,v_{\hat\rho}\,\phi_1 + u_1\,f''(u_0)\,v\,\phi_1 + g'(u_0)\,v\,\phi_1\\
&\qquad\qquad \left.+ \frac{\hat\rho}{\hat R}\,u_1\,f''(u_0)\,v\,\phi_0 + \frac{\hat\rho}{\hat R}\,g'(u_0)\,v\,\phi_0\Bigg]\right\}\,d\hat\rho\,d\hat s
\end{aligned}\tag{34}$$

since $v = w(\hat s)u_0^{'}(\hat\rho)$ is in the kernel of L.

Equation (34) has terms involving u_1 and u_2, so we must examine (16) and (17) for these terms. We take the derivative of (16) with respect to $\hat\rho$ and multiply by $w\phi_0$, integrate, and use Green's identity to obtain

$$\int_{\partial\Omega}\partial_n u_1' w\phi_0\,ds = \int_\Omega\left(-\frac{1}{\hat R}u_0'' - f''(u_0)u_0'u_1 - g'(u_0)u_0'\right)w\phi_0\,dA\,. \tag{35}$$

It will become evident that $\lambda = O(\varepsilon^2)$. To avoid tedious calculations, we will write $\lambda = \varepsilon^2\lambda_0 + \cdots$. In this way, λ_0 terms will enter at the correct order. We substitute (31) and (32) into (35), multiply by ε, and arrange the terms to match the u_1 term in (34):

$$\begin{aligned}
&\varepsilon^2\int_\Omega f''(u_0)u_0'u_1 w\phi_0\,d\hat\rho\,d\hat s\\
&\qquad = -\,\varepsilon^2\left(\int_\Omega\left(\frac{1}{\hat R}u_0'' + g'(u_0)\,u_0'\right)w\phi_0\,d\hat\rho\,d\hat s + \int_{\partial\Omega}\partial_n u_1' w\phi_0\,dt\right)\\
&\qquad\quad + \varepsilon^3\int_\Omega\left(-\frac{1}{\hat R}u_0'' - f''(u_0)u_0'u_1 - g'(u_0)\,u_0'\right)\frac{\hat\rho}{\hat R}w\phi_0\,d\hat\rho\,d\hat s + \cdots\,.
\end{aligned}\tag{36}$$

We repeat the above procedure to handle the u_2 term in (34). First we differentiate (17) with respect to $\hat\rho$,

$$\begin{aligned}
\Delta(u_2') + f'(u_0)u_2' = {}& -\frac{1}{\hat R}u_1'' + \frac{1}{\hat R^2}\hat\rho u_0'' + \frac{1}{\hat R^2}u_0' - f''(u_0)u_0'u_2\\
&-\ \frac{f'''(u_0)u_0'u_1^2}{2} - f''(u_0)u_1u_1' - g''(u_0)u_0'u_1 - g'(u_0)u_1'\,.
\end{aligned}\tag{37}$$

We multiply the above expression by ϕ_0, integrate over the domain, apply Green's identity to the right-hand side, and multiply by ε^3 to match the u_2 term in (34) which results in the following:

$$
\begin{aligned}
\varepsilon^3\!\int_\Omega f''(u_0)u_0'u_2w\phi_0\,d\hat\rho\,d\hat s = \varepsilon^3\!\int_\Omega \Bigg(&-\frac{1}{\hat R}u_1''+\frac{1}{\hat R^2}\hat\rho u_0''+\frac{1}{\hat R^2}u_0'-\frac{f'''(u_0)u_0'u_1^2}{2}\\
&-f''(u_0)u_1u_1'-g''(u_0)u_0'u_1-g'(u_0)u_1'\Bigg)w\phi_0\,d\hat\rho\,d\hat s\\
&+\varepsilon^3\!\int_{\partial\Omega}\partial_n u_2' w\phi_0\,dt+\cdots\ . \qquad (38)
\end{aligned}
$$

Since ϕ is a translation eigenfunction, in the interior we may write

$$\phi_i = T(\hat s)u_i'(\hat\rho). \qquad (39)$$

We also note that

$$\int_\Omega \frac{1}{\hat R}u_1''w\phi_0\,d\hat\rho\,d\hat s = -\int_\Omega\frac{1}{\hat R}w\phi_0'u_1'\,d\hat\rho\,d\hat s\,. \qquad (40)$$

Using (40), (39), (38), (36), and (34) we can write (28) as

$$
\begin{aligned}
(41)\quad &\varepsilon^2\lambda_0\int_\Omega v\phi_0\,d\hat\rho\,d\hat s\\
&=\varepsilon^2\!\int_\Omega\left(v_{\hat s\hat s}\phi_0+\frac{2}{\hat R}\phi_1 v_{\hat\rho}+\frac{1}{\hat R^2}w\phi_0\right)d\hat\rho\,d\hat s-\varepsilon^2\!\int_{\partial\Omega}(\phi_0\partial_n v+w\phi_0\partial_n u_1')\,dt+\cdots,
\end{aligned}
$$

where, from (38), the boundary integral involving $\partial_n u_2'$ is of higher order. The eigenfunction $\phi_0 = T(\hat s)u_0'$ is the derivative of a monotonic front and is, thus, of one sign and hence is the principal eigenfunction. The principal eigenfunction of L must be even in the radial direction and the function v' will be odd in the radial direction. Thus, the term $\int_\Omega \frac{2}{\hat R}\phi_1 v'\,d\hat\rho\,d\hat s$ will be zero to leading order.

For the boundary integral involving $\partial_n v$, we need to find $\partial_n v$ on $\partial\Omega$. Away from the points where the front and boundary intersect, $\partial_n v$ will be exponentially small, so we will only consider the two components of the boundary $\Gamma_\pm$. Since the front meets $\Gamma_\pm$ orthogonally,

$$\partial_n v|_{\Gamma_\pm} = \left.\frac{\partial v}{\partial r}\right|_{\Gamma_\pm}. \qquad (42)$$

We note from (8), (9), (10), and (11),

$$\left.\frac{\partial\hat s}{\partial r}\right|_{\Gamma_\pm}\sim\pm1 \quad\text{and}\quad \left.\frac{\partial\hat\rho}{\partial r}\right|_{\Gamma_\pm}\sim\left.\frac{t}{R}\right|_{\Gamma_\pm}. \qquad (43)$$

Thus,

$$\partial_n v\big|_{\Gamma_\pm}\sim\left.\left(\pm w'(\hat s)u_0'(t)+w(\hat s)u_0''(t)\frac{t}{R}\right)\right|_{\Gamma_\pm}. \qquad (44)$$

We let ℓ be the length of the interface and place $\hat s = 0$ such that $\hat s = \pm\ell/2$ on $\Gamma_\pm$. Then, using (39), (29), $\hat\rho\sim t$ on $\Gamma_\pm$ and $\int tu_0''u_0 = -\frac{1}{2}\int {u_0'}^2$ with (44) results in

$$
\begin{aligned}
-\int_{\partial\Omega}\partial_n v\phi_0\,dt &\sim w'(-\ell/2)T(-\ell/2)\int_{-\infty}^{\infty}(u_0'(t))^2\,dt+\frac{w(-\ell/2)T(-\ell/2)}{2R_-}\int_{-\infty}^{\infty}(u_0'(t))^2\,dt\\
&\quad-w'(\ell/2)T(\ell/2)\int_{-\infty}^{\infty}(u_0'(t))^2\,dt+\frac{w(\ell/2)T(\ell/2)}{2R_+}\int_{-\infty}^{\infty}(u_0'(t))^2\,dt\,. \qquad (45)
\end{aligned}
$$

For the boundary integral involving $\partial_n u_1'$, we have that, near $\partial\Omega$,

$$u \sim u_0(\hat\rho) + \varepsilon\, u_1 = u_0(t) + \varepsilon \left(\frac{\rho\, t}{R} + \frac{\rho^2}{2\hat R} \right) u_0'(t) + \varepsilon\, u_1 + \cdots . \tag{46}$$

Also, on $\partial\Omega$, we have $\partial_n u = 0$, so that, on $\partial\Omega$

$$\partial_n u_1 \sim - \frac{\partial}{\partial\rho} \left[\frac{1}{\varepsilon}\, u_0(t) + \left(\frac{\rho\, t}{R} + \frac{\rho^2}{2\hat R} \right) u_0'(t) \right] \Bigg|_{\rho=0} = - \frac{t}{R}\, u_0'(t) \Bigg|_{\Gamma_\pm} . \tag{47}$$

Then

$$\partial_n u_1' \sim - \frac{1}{R}\, u_0'(t) - \frac{t}{R}\, u_0''(t) \tag{48}$$

and

$$\begin{aligned} - \int_{\partial\Omega} \partial_n u_1' w \phi_0 \, dt \sim & \int_{\Gamma_-} w(-\ell/2) T(-\ell/2) \left(u_0''(t) u'(t) \frac{t}{R} + u_0'(t)^2 \frac{1}{R} \right) dt \\ & + \int_{\Gamma_+} w(\ell/2) T(\ell/2) \left(u_0''(t) u'(t) \frac{t}{R} + u_0'(t)^2 \frac{1}{R} \right) dt \,, \\ = & \left(\frac{w(\ell/2)T(\ell/2)}{2R_+} + \frac{w(-\ell/2)T(-\ell/2)}{2R_-} \right) \int_{-\infty}^{\infty} (u_0'(t))^2 \, dt \,. \end{aligned} \tag{49}$$

Substitute (44) and (49) into (41) to obtain

$$\begin{aligned} \left(\lambda_0 - \frac{1}{\hat R^2} \right) \int_\Omega v \phi_0 \, d\hat\rho \, d\hat s \sim \int_\Omega v_{\hat s \hat s} \phi_0 \, d\hat\rho \, d\hat s + & \left(w'(-\ell/2)T(-\ell/2) + \frac{w(-\ell/2)T(-\ell/2)}{2R_-} \right. \\ & \left. - w'(\ell/2)T(\ell/2) + \frac{w(\ell/2)T(\ell/2)}{2R_+} \right) \int_{-\infty}^{\infty} (u_0'(t))^2 \, dt \,. \end{aligned} \tag{50}$$

The eigenfunctions will depend on both $\hat s$ and $\hat\rho$. We thus substitute the ansatz $\phi = T(\hat s)\Phi(\hat\rho)$ into the eigenvalue problem (4),

$$\left(\Phi'' + \frac{\varepsilon}{\hat R} \Phi' - \varepsilon^2 \frac{\hat\rho}{\hat R^2} \Phi' + f'(u)\Phi + \varepsilon g'(u)\Phi \right) T + \varepsilon^2 T'' \Phi = \varepsilon^2 \lambda_0 T \Phi \,. \tag{51}$$

We divide both sides by $T\Phi$,

$$\left(\frac{\Phi'' + \frac{\varepsilon}{\hat R}\Phi' - \varepsilon^2 \frac{\hat\rho}{\hat R^2}\Phi' + f'(u)\Phi + \varepsilon g'(u)\Phi}{\Phi} \right) + \varepsilon^2 \frac{T''}{T} = \varepsilon^2 \lambda_0 \,. \tag{52}$$

Since T is independent of $\hat\rho$, the term in the brackets must be independent of $\hat\rho$ or equal to a constant α:

$$\Phi'' + \frac{\varepsilon}{\hat R} \Phi' - \varepsilon^2 \frac{\hat\rho}{\hat R^2} \Phi' + f'(u)\Phi + \varepsilon g'(u)\Phi = \alpha \Phi \,. \tag{53}$$

We expand $\Phi = \Phi_0 + \varepsilon\Phi_1 + \varepsilon^2\Phi_2 + \cdots$ and $\alpha = \alpha_0 + \varepsilon\alpha_1 + \varepsilon\alpha_2 + \cdots$. The lowest order terms satisfy

$$\Phi_0'' + f'(u_0)\Phi_0 = \alpha_0 \Phi_0 \,. \tag{54}$$

Thus, $\Phi_0 = u_0'(\hat\rho)$ and $\alpha_0 = 0$. The $O(\varepsilon)$ terms satisfy

$$\Phi_1'' + f'(u_0)\Phi_1 = \alpha_1\Phi_0 - \frac{1}{\hat R}\Phi_0' - f''(u_0)u_1\Phi_0 - g'(u_0)\Phi_0\,. \tag{55}$$

Differentiating (16) results in the following solvability condition,

$$\int_{-\infty}^{\infty} f''(u_0)u_1(u_0')^2\,d\hat\rho = -\int_{-\infty}^{\infty} g'(u_0)(u_0')^2\,d\hat\rho\,. \tag{56}$$

Applying (56) to the solvability condition for (55) yields $\alpha_1 = 0$. The $O(\varepsilon^2)$ terms satisfy

$$\begin{aligned}\Phi_2'' + f'(u_0)\Phi_2 &= \alpha_2\Phi_0 - \Phi_1'\frac{1}{\hat R} + \frac{1}{\hat R^2}\hat\rho\Phi_0' \\ &\quad - f''(u_0)u_1\Phi_1 - f''(u_0)u_2\Phi_0 - \frac{1}{2}f'''(u_0)u_1^2\Phi_0 - g''(u_0)u_1\Phi_0 - g'(u_0)\Phi_1\,.\end{aligned} \tag{57}$$

We have the following solvability condition:

$$\begin{aligned}\alpha_2\int_{-\infty}^{\infty}\Phi_0^2\,d\hat\rho =& \frac{1}{2}\int_{-\infty}^{\infty} f'''(u_0)u_1^2\Phi_0^2\,d\hat\rho + \int_{-\infty}^{\infty} g''(u_0)u_1\Phi_0^2\,d\hat\rho + \int_{-\infty}^{\infty} g'(u_0)\Phi_1\Phi_0\,d\hat\rho \\ &+ \int_{-\infty}^{\infty} f''(u_0)u_2\Phi_0^2\,d\hat\rho + \int_{-\infty}^{\infty} f''(u_0)u_1\Phi_1\Phi_0\,d\hat\rho - \int_{-\infty}^{\infty}\frac{1}{\hat R^2}\hat\rho\Phi_0'\Phi_0\,d\hat\rho \\ &+ \int_{-\infty}^{\infty}\frac{1}{\hat R}\Phi_1'\Phi_0\,d\hat\rho\,.\end{aligned} \tag{58}$$

Differentiating (17) results in the solvability condition,

$$\begin{aligned}&-\int_{-\infty}^{\infty} f''(u_0)u_2(u_0')^2\,d\hat\rho - \int_{-\infty}^{\infty}\frac{1}{\hat R}u_1''u_0'\,d\hat\rho + \frac{1}{\hat R^2}\int_{-\infty}^{\infty}\hat\rho u_0''u_0'\,d\hat\rho + \frac{1}{\hat R^2}\int_{-\infty}^{\infty}(u_0')^2\,d\hat\rho \\ &-\int_{-\infty}^{\infty} f''(u_0)u_1u_1'u_0'\,d\hat\rho - \frac{1}{2}\int_{-\infty}^{\infty} f'''(u_0)u_1^2(u_0')^2\,d\hat\rho - \int_{-\infty}^{\infty} g''(u_0)u_1(u_0')^2\,d\hat\rho \\ &-\int_{-\infty}^{\infty} g'(u_0)u_1'u_0'\,d\hat\rho = 0\,.\end{aligned} \tag{59}$$

Now we use $\int_{-\infty}^{\infty}\hat\rho u_0''u_0'\,d\hat\rho = -\frac{1}{2}\int_{-\infty}^{\infty}(u')^2\,d\hat\rho$ and (58) in (57) to yield

$$\alpha_2 = \frac{1}{\hat R^2}\,. \tag{60}$$

Now we can substitute (53) into (52) using $\alpha = \frac{\varepsilon^2}{\hat R^2} + \cdots$ to get

$$T'' = \left(\lambda_0 - \frac{1}{\hat R^2}\right)T\,. \tag{61}$$

Note that

$$\left(\lambda_0 - \frac{1}{\hat{R}^2}\right) \int_\Omega v\phi_0 \, d\hat{\rho} \, d\hat{s} \sim \left(\lambda_0 - \frac{1}{\hat{R}^2}\right) \int_{-\ell/2}^{\ell/2} wT d\hat{s} \int_{-\infty}^{\infty} (u_0'(t))^2 \, dt \tag{62}$$

$$\int_\Omega v_{\hat{s}\hat{s}}\phi_0 \sim \int_{-\ell/2}^{\ell/2} w^{''} T d\hat{s} \int_{-\infty}^{\infty} (u_0'(t))^2 \, dt \, . \tag{63}$$

Substituting (61), (62), and (63) into (50), integrating by parts, we obtain

$$w(-\ell/2)\left[T^{'}(-\ell/2) + \frac{1}{R_-}T(-\ell/2)\right] + w(\ell/2)\left[-T^{'}(\ell/2) + \frac{1}{R_+}T(\ell/2)\right] = 0.$$

Since w is an arbitrary test function, we see that T satisfies the following boundary conditions:

$$T^{'}(-\ell/2) + \frac{1}{R_-}T(-\ell/2) = 0, \;\; -T^{'}(\ell/2) + \frac{1}{R_+}T(\ell/2) = 0. \tag{64}$$

Equations (61) and (64) prove that T satisfies the geometric eigenvalue problem (5b). Hence, $\lambda_0 = \frac{1}{\hat{R}^2} - \alpha$ where α satisfies

$$\begin{cases} T'' + \alpha T = 0, \\ T'(-\ell/2) + \kappa_- T(-\ell/2) = 0, \\ T'(\ell/2) - \kappa_+ T(\ell/2) = 0, \end{cases} \tag{65}$$

where $\kappa_\pm \equiv \frac{1}{R_\pm}$ and $\kappa_\pm > 0$ corresponds to a convex domain as in Figure 3.

If $\alpha \leq 0$, then $\lambda_0 \geq \frac{1}{\hat{R}^2}$. If $\alpha = \mu^2 > 0$, (where $\mu > 0$), then it is easy to see that μ must satisfy the following transcendental relation:

$$\tan(\mu\ell) = \frac{\mu(\kappa_+ + \kappa_-)}{\kappa_+\kappa_- - \mu^2} \, , \tag{66}$$

and the eigenvalues of (4) are given by

$$\varepsilon^2\lambda = \frac{1}{\hat{R}} - \mu^2 \, , \tag{67}$$

which is precisely (6). Formula (7) is seen to be identical to (6) by applying the identity

$$\tan(x+y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}. \tag{68}$$

This completes the proof of Theorem 2. □

4. Proof of Theorem 3. In this section we show that the geometric condition of Theorem 3 is a direct consequence of Theorem 2.

Fix a point q_+ on the top boundary and consider a circular arc going through q_+ and intersecting both top and bottom boundaries orthogonally (refer to Figure 4). Let p be the center of this arc and let R denote its radius. First, we shall show that $\frac{dR}{dq_+} = 0$ if and only if the formula (5) holds with $\lambda_0 = 0$. By zooming into the point where $\frac{dR}{dq_+} = 0$, we can assume that locally, p moves along a straight line as q_+ moves along the boundary, and that the boundaries are segments of circles of radii $\mathcal{R}_\pm$, as

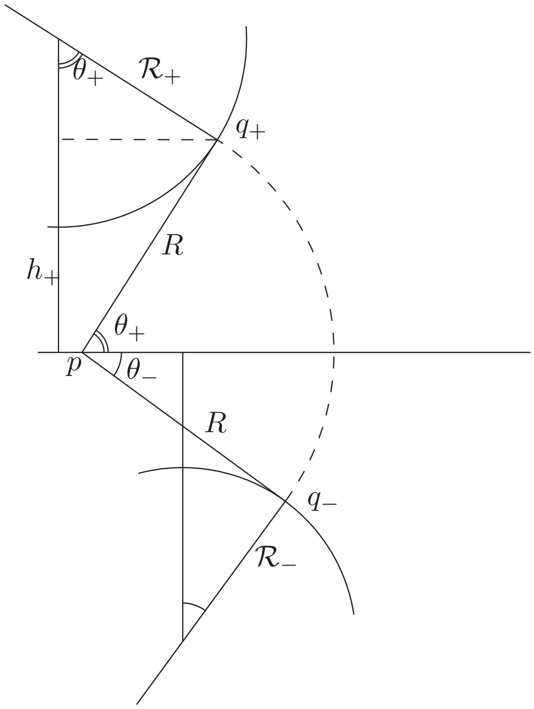


FIG. 4. *Setup for proof of Theorem* 3.

shown in Figure 4. In general, $\mathcal{R}_\pm$ may be positive or negative; for convenience, as shown in the figure, we chose $\mathcal{R}_\pm = -\frac{1}{\kappa_\pm}$ with $\kappa_\pm < 0$ so that $\mathcal{R}_\pm$ is positive. Now from geometry, we find the relationship

$$R = \frac{\mathcal{R}_+(1-\cos\theta_+) + h_+}{\sin\theta_+},$$

where h_+, θ_+ are as shown in Figure 4. We obtain

$$\frac{\partial R}{\partial \theta_+} = \frac{\mathcal{R}_+ - (\mathcal{R}_+ + h_+)\cos\theta_+}{\sin^2\theta_+}$$

so that upon eliminating h_+ we obtain

$$\frac{\partial R}{\partial \theta_+} = 0 \iff \frac{R}{\mathcal{R}_+} = \tan\theta_+ \tag{69}$$

and similarly with + replaced by −. Since $\theta_\pm$ are functions of q_+, we find that at the point where $\frac{dR}{dq_+} = 0$, we have

$$\arctan\frac{R}{\mathcal{R}_\pm} = \theta_\pm.$$

Now from geometry, $\theta_+ = \ell_+/R$, $\theta_- = \ell_-/R$, and $\ell = \ell_+ + \ell_-$. Therefore, upon adding the two equations in (69) we obtain

$$\arctan\frac{R}{\mathcal{R}_+} + \arctan\frac{R}{\mathcal{R}_-} = \theta_+ + \theta_- = \frac{\ell}{R}.$$

But this is precisely (7) with $\lambda_0 = 0$ after substituting $\mathcal{R}_\pm = -\frac{1}{\kappa_\pm}$.

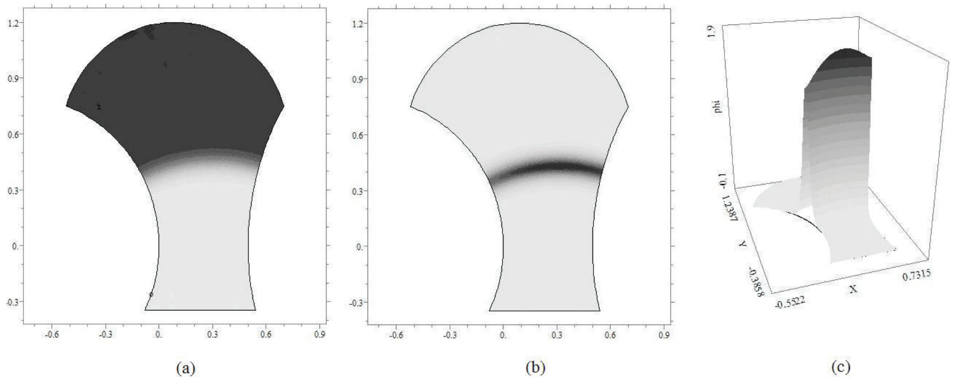


FIG. 5. *Numerical computation of interface and eigenvalue. Left: the steady-state solution $u(x)$ of* (70). *Dark denotes $u \sim 1$ and light denotes $u \sim -1$. Middle: The shape of the corresponding eigenfunction ϕ. Right: surface plot of ϕ. Note the sinusoidal shape along the direction of the interface boundary. Note also a corner layer that is evident near the boundary of the domain. See section* 5 *for parameter values.*

Next, we note that in the case of a cone ($\kappa_+ = \kappa_- = 0$), (7) yields $\lambda_0 = \frac{1}{R^2} > 0$ so that the interface is unstable for a cone domain, for which $R' > 0$. Since λ_0 can only be real, it follows by continuity that λ_0 crosses zero if and only if $R' = 0$, and λ_0 is negative if and only if $R' < 0$. This concludes the proof. □

5. Numerical example.

We now provide a numerical example of Theorem 2. All computations were done using using the software FlexPDE [19].

Consider a domain as shown in Figure 5. Its left and right boundaries consist of arcs of circles of radii $\mathcal{R}_- = 0.8, \ \ \mathcal{R}_+ = 1.5$, so that $\kappa_- = -1.25, \ \ \kappa_+ = -0.667$. The distance between these two boundaries was chosen to be 0.5. The shape of the top and bottom boundaries does not affect the computation as long as they are located $O(1)$ distance from the interface. We chose the nonlinearity to be

$$u_t = \varepsilon^2 \Delta u - 2(u - \varepsilon a)(u-1)(u+1) \tag{70}$$

with $a = 0.55, \ \ \varepsilon = 0.06$. From Proposition 1 we obtain the theoretical value of the interface radius to be $\hat{R} = \frac{1}{2a} = 0.9091$. To estimate the numerical value of $\hat{R}$, we have used FlexPDE to compute the steady state solution to (70), using $u = \tanh(y/\varepsilon)$ as initial conditions. The resulting steady state is shown on Figure 5(a). Next, we computed the coordinates of the intersection of the middle of the interface ($u = 0$) with the boundary, and then used geometry to obtain $\hat{R}_{\text{numerical}} = 0.9066$. This is in excellent agreement with the theoretical prediction. Geometry then yields an estimate of $l = 0.6486$.

Next, we have solved the eigenvalue problem (4) numerically. Using a global error tolerance of 0.5×10^{-4}, we obtained a numerical estimate of $\lambda_{\text{numerical}} = 0.00504$. This required about 10,000 gridpoints (FlexPDE uses adaptive gridding, and chooses the mesh size based on the global tolerance setting. We have also verified that this result is correct to two significant digits by changing the tolerance). On the other hand, solving (6) gives the theoretical estimate of $\lambda = 0.00506$. Excellent agreement (within 0.5%) is observed.

6. Discussion.

In this paper we have characterized the stability of curved interfaces of the perturbed AC system on a bounded domain. On one hand, it is a

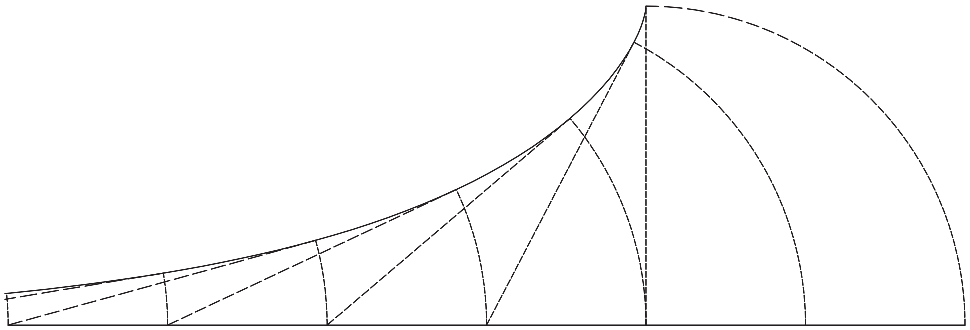


FIG. 6. *A tractrix: the threshold case where all circles intersecting the boundary have identical radius. Theorem* 3 *does not apply to such a domain.*

generalization of the geometric eigenvalue problem derived in [10], [11] for the Allen–Cahn equation without perturbations, which only admits straight interfaces. On the other hand, the perturbed system (1) has been studied on the whole space $\mathbb{R}^2$ without the boundaries—see, for example, [16], [15]. We show that the presence of *both* boundaries and a perturbation can stabilize a curved front. By contrast, the curved front is always unstable in the absence of boundaries—it either shrinks to a point or expands indefinitely depending on the initial conditions [15]. To our knowledge, the characterization of stability that combines both the curvature of the interface and the boundary effect is new.

Algebraically, the stability condition is given by Theorem 2. Geometrically, Theorem 3 states that if $R(s)$ denotes the radius of an arc that intersects the boundary orthogonally at $q_{\pm}(s)$, then the interface is stable if $R'(s) < 0$ whenever $R = \hat{R}$, whereas the interface is unstable if $R'(s) > 0$ at that point (see Figure 2). In particular, this shows explicitly the well-known result that an interface at equilibrium cannot be stable in a convex domain [3]; on the other hand we have shown numerical and theoretical examples where such interface is stable when the domain is nonconvex.

In general, the relationship between the radius R of a circle that intersects the boundary orthogonally and the domain boundary $q = (x, y)$ is given by

$$x = p_1 + R\cos\theta, \qquad y = p_2 + R\sin\theta,$$

where $p = (p_1, p_2)$ is the center of the arc of radius R; p_1, p_2, R are arbitrary functions of s; and θ satisfies a differential equation

$$R\frac{d\theta}{ds} = p_1' \sin\theta - p_2' \cos\theta.$$

An interesting threshold case corresponds to $R = \hat{R}$ for all s. If the bottom boundary is the x-axis and $R = \hat{R}$ for all s, then the top boundary forms a *tractrix* (see Figure 6.) This is a well-known curve that is also generated when a ball is dragged on a fixed string by a tractor moving along the x-axis. Implicitly, this curve is given by

$$x = \hat{R}(-t + \tanh(t)), \quad y = \hat{R}\,\mathrm{sech}(t).$$

It is an open problem to describe either the stability or the location of the interface for such a domain.

An interesting conjecture arises in studying the propagation of fronts around a concave corner. Such domains were used in [17], where the propagation of chemical fronts was considered. An interface passing through the corner may get “stuck” at the corner or go through it, depending on the geometry. If we “smooth out” the corner and take ε sufficiently small, then we can apply Theorem 3. The result is that the interface will get stuck at the corner if there exists a circle that intersects orthogonally with one boundary, and that passes through the corner point, and whose radius is at most $\hat{R}$. This is essentially the geometrical condition described in section III.B in [17] and it agrees well with numerical results presented there. However, the construction of an interface at a corner point is an open theoretical problem.

REFERENCES

[1] N. Alikakos, G. Fusco, and M. Kowalczyk, *Finite dimensional dynamics and interfaces intersecting the boundary: Equilibria and the quasi-invariant manifold*, Indiana Univ. Math. J., 45 (1996), pp. 1119–1156.

[2] S. Allen and J. W. Cahn, *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta. Metall., 27 (1979), pp. 1084–1095.

[3] R. G. Casten and J. Holland, *Instability results for reaction diffusion equations with Neumann boundary conditions*, J. Differential Equations, 27 (1978), pp. 266–273.

[4] G. Chapuisat and E. Grenier, *Existence and nonexistence of traveling wave solutions for a bistable reaction-diffusion equation in an infinite cylinder whose diameter is suddenly increased*, Comm. Partial Differential Equations, 30 (2005), pp. 1805–1816.

[5] X. Chen, *Generation and propagation of inerfaces in reaction-diffusion equations*, J. Differential Equations, 96 (1992), pp. 116–141.

[6] M. del Pino, M. Kowalczyk, and J. Wei, *Resonance and interior layers in an inhomogeneous phase transition model,* SIAM J. Math. Anal., 38 (2007), pp. 1542–1564.

[7] P. C. Fife and J. B. MacLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch Ration. Mech. Anal., 65 (1977), pp. 335–361.

[8] P. Fife and W. M. Greenlee, *Interior transition layers for elliptic boundary value problems with a small parameter,* Russian Math. Surveys, 29 (1974), pp. 103–131.

[9] T. Ilmanen, *Convergence of the Allen-Cahn equation to the Brakkke's motion by mean curvature*, J. Differential Geom., 38 (1993), pp. 417–461.

[10] M. Kowalczyk, *Approximate invariant manifold of the Allen-Cahn flow in two dimensions*, in Partial Differential Equations and Inverse Problems, Comptemp. Math. 362, Amer. Math. Soc., Providence, RI, 2004, pp. 233–239.

[11] M. Kowalczyk, *On the existence and Morse index of solutions to the Allen-Cahn equation in two dimensions*, Ann. Mat. Pura Appl., 184 (2005), pp. 0373–3114.

[12] R. Kohn and P. Sternberg, *Local minimizers and singular perturbations*, Proc. R. Soc. Edinburgh, 111 A (1989), pp. 69–84.

[13] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. Res. Inst. Math. Sci., 15 (1979), pp. 401–454.

[14] F. Pacard and M. Ritore, *From constant mean curvature hypersurfaces to the gradient theory of phase transition*, J. Differential Geom., 64 (2003), pp. 359–423.

[15] L. Pismen, *Patterns and Interfaces in Dissipative Dynamics,* Springer Verlag, 2006, p. 374.

[16] J. Rubinstein, P. Sternberg, and J. B. Keller, *Fast reaction, slow diffusion, and curve shortening*, SIAM J. Appl. Math., 49 (1989), pp. 116–133.

[17] L. Qiao, I. G. Kevrekidis, C. Punckt, and H. H. Rotermund, *Guiding chemical pulses through geometry: Y junctions*, Phys. Rev. E, 73 (2006), 036219.

[18] M. Ward and D. Stafford, *Metastable dynamics and spatially inhomogeneous equilibria in dumbbell-shaped domains*, Studies in Appl. Math., 103 (1999), pp. 51–73.

[19] See FlexPDE website, www.pdesolutions.com.

SIAM J. Appl. Math.
Vol. 69, No. 5, pp. 1244–1262

THREE LIMIT CYCLES IN A LESLIE–GOWER PREDATOR-PREY MODEL WITH ADDITIVE ALLEE EFFECT*

PABLO AGUIRRE†, EDUARDO GONZÁLEZ-OLIVARES‡, AND EDUARDO SÁEZ†

Abstract. In this work, a bidimensional continuous-time differential equations system is analyzed which is derived of Leslie-type predator-prey schemes by considering a nonmonotonic functional response and Allee effect on population prey. For the system obtained we describe the bifurcation diagram of limit cycles that appears in the first quadrant, the only quadrant of interest for the sake of realism. We show that, under certain conditions over the parameters, the system allows the existence of three limit cycles: The first two cycles are infinitesimal ones generated by Hopf bifurcation; the third one arises from a homoclinic bifurcation. Furthermore, we give conditions over the parameters such that the model allows long-term extinction or survival of both populations. In particular, the presence of a weak Allee effect does not imply extinction of populations necessarily for our model.

Key words. stability, limit cycles, homoclinic orbits, bifurcations, predator-prey models, Allee effect

AMS subject classifications. 92D25, 34C, 58F14, 58F21

DOI. 10.1137/070705210

1. Introduction.

This work deals with a continuous predator-prey model considering the following: (i) the Allee effect [6, 13, 16, 34] affecting the prey population, (ii) the functional response of predators of nonmonotonic type, and (iii) a predator growth function of logistic type. Other inherent assumptions for the model are that population size varies only in time and it is uniformly distributed in space, there is no division of ages or sex, and it is not affected by abiotic factors.

It is known that the Allee effect refers to a positive density dependence in prey population growth at low prey densities [34], and it occurs whenever fitness of an individual in a small or sparse population decreases as the population size or density also declines [6]. To be able to recognize the consequences of the Allee effect on reproduction, conservation and behavior of species has become an important aim over the last years. The analysis and understanding of this phenomenon can bring important benefits not only for ecology but also for various applied engineering disciplines such as agropecuary, fishing, and forestal industries.

Different mechanisms generating Allee effects have been suggested (see Table 1 in [6]), being largely studied as singular entities and usually describing a situation in which the population growth rate decreases under some minimum critical density [5] or when a limited population growth capacity is observed [15]. In some cases this growth rate might be even negative, causing an extinction threshold [5]. In other words, the Allee effect may be understood as the cause of the increase in extinction risk at low densities [16], introducing in some cases a population threshold that has to be exceeded by population to be able to grow. This effect is also named in population dynamics as the *negative competition effect* [38]; in fisheries sciences, it is called *depensation*

*Received by the editors October 12, 2007; accepted for publication (in revised form) October 6, 2008; published electronically February 4, 2009. This work was partially supported by USM grant 12.08.04.

http://www.siam.org/journals/siap/69-5/70521.html

†Departamento de Matemática, Universidad Técnica Federico Santa María, Casilla 110-V, Valparaíso, Chile (pablo.aguirre@usm.cl, eduardo.saez@usm.cl).

‡Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile (ejgonzal@ucv.cl).

[9, 16], and in epidemiology, its analogous is the *eradication threshold*, the population level of susceptible individuals below which an infectious illness is eliminated from a population [5].

More precisely, the (component) Allee effect is any positive relationship between any measurable component of individual fitness and population size or density [6, 13, 34], and it might be the *strong Allee effect* [6] or *critical depensation* [9] that implies the existence of a threshold population level [5, 9], whereas the *weak Allee effect* [34] or *noncritical depensation* [9] does not have it. When the species is submitted to a strong Allee effect, it may have a bigger tendency to be less able to overcome these additional mortality causes, to have a slower recovery, and to be more prone to extinction than other species [34].

In most predation models it has been considered that the Allee effect has influence only on the prey population, and this effect is independent of the functional response or consumption rate that reflects the change on predation due to the prey's population size. Quantitatively, it is assumed that the functional response has influence on the extension of the bistability region [15].

The Allee effect has been modeled in different ways using various mathematical tools and, in a first approach, like a deterministic phenomenon frequently associated to population's stochastic fluctuations [5]. For instance, if $x = x\left(t\right)$ indicates the population size, the most usual continuous growth equation to express the Allee effect is given by

$$\frac{dx}{dt} = r\ \left(1 - \frac{x}{K}\right)\left(x - m\right)x,$$

featuring the *multiplicative Allee effect.* Clearly, if $m = 0$, we have the weak Allee effect and if $m > 0$, it has the strong Allee effect.

Other mathematical forms have been proposed to describe this phenomenon [8]. In this work, we consider the natural growth function deduced in [16] and [34] given by the equation

$$\frac{dx}{dt} = r\ \left(1 - \frac{x}{K} - \frac{m}{x+b}\right)x \tag{1}$$

that we call *additive Allee effect.*

Recent ecological research suggests the possibility that two or more Allee effects generate mechanisms acting simultaneously on a single population (see Table 2 in [6]), and the combined influence of some of these phenomena has been named the *multiple Allee effects.*

In the interacting populations, the predation can be largely reduced due to better ability of prey to avoid predation when their population size is large enough [42, 43]; but at low population densities, there could be a low effectiveness of antipredator vigilance, which reflects an Allee effect. For some marine species it has been shown that the per capita growth decreases as the size population is reduced below some critical level, and two proposed causes of depensation or Allee effect are the following: reduced breedings success at low densities of the population and increased relative predation on small populations [21]. This situation has happened in many real fisheries as a result of overfishing when man acts as predator [9].

Considering these aspects, (1) can be rewritten as

$$\frac{dx}{dt} = \frac{r\ x}{x+b}\ \left(1 - \frac{x}{x_K}\right)(x - x_m)x, \tag{2}$$

where $x_m = \frac{1}{2}(K - b - \sqrt{\frac{1}{r}(r(K+b)^2 - 4mK)})$ and $x_K = \frac{1}{2}(K - b + \sqrt{\frac{1}{r}(r(K+b)^2 - 4mK)})$ for $r(K+b)^2 - 4mK > 0$. Hence, (2) represents two types of Allee effect affecting the same population since x_m expresses the minimum of viable population and the factor $r\,(x) = \frac{r\,x}{x+b}$ indicates the impact of an Allee effect due to other causes affecting the intrinsic growth rate, for example, the predation reducing breeding success at low densities [9, 21].

On the other hand, the *functional response of predators* or *consumption rate function* refers to the change in the density of prey attached per unit time per predator as the prey density changes [20, 42]. In most predator-prey models considered in the ecological literature, the predator response to prey density is assumed to be monotonic increasing, the inherent assumption being that as there is more prey in the environment, it is better for the predator [20].

However, there is evidence that indicates that this need not always be the case, for instance, when a type of antipredator behavior (APB) exists. *Group defense* is one of these, and the term is used to describe the phenomenon whereby predators decrease, or are even prevented altogether, due to the increased ability of the prey to better defend or disguise themselves when their number is large enough [20, 42, 39, 43], and in this case a nonmonotonic functional response is better. For example, lone musk ox can be successfully attacked by wolves; however, large herds of them can be attacked but with rare success.

Another manifestation of an APB in which a nonmonotonic functional response (or Holling-type IV or Monod Haldane) can be used is the phenomenon of *aggregation*, a social behavior of prey in which prey congregate on a fine scale relative to the predator so that the predator's hunting is not spatially homogeneous [36], such as succeeds with mile-long schools of certain class of fishes. In this case, a primary advantage of schooling seems to be confusion of the predator when it attacks. The more important benefit of aggregation is an increasing in wariness. Moreover, aggregation can both decrease the vulnerability to be attacked and increase the time that group members can devote to activities other than surveillance [36].

Other related examples of nonmonotone consumption occur at the microbial level where evidence indicates that when faced with an overabundance of nutrient the effectiveness of the consumer can begin to decline. This is often seen when microorganisms are used for waste decomposition or for water purification, a phenomenon that is called *inhibition* [20, 42, 43].

In these cases, the functional response curves have an upper bound on the rate of predation per individual predator at some prey density, in contrast to the old Lotka–Volterra model which assumed a linear relationship between prey density and the rate of predation over the entire range of prey densities [36]. In this work, we use the function $h(x) = \frac{qx}{x^2+a}$, also employed in [28, 42, 39, 43] and corresponding to the Holling-type IV functional response [36], which is generalized as $h(x) = \frac{qx}{x^2+bx+a}$ in [44, 40] for a Gause model. This generalized expression is derived by Collings [12], who affirms that this type of functional response seems a reasonable possibility if it is assumed that prey and webbing densities are directly related. In [40], for the corresponding Gause model the existence of two limit cycles is proven. Moreover, this Gause model exhibits bifurcation of cusp type with codimension two [4, 23] or Bogdanov–Takens bifurcation [44, 40].

We note that the phenomena of the Allee effect and aggregation described by a nonmonotonic functional response are quite compatible and justify our assumptions in the model studied.

The last aspect in our equations is a feature of Leslie-type predator-prey models [37] or the Leslie–Gower model [27], in which the conventional environmental carrying capacity K_y is proportional to prey abundance x [31], that is, $K_y = nx$, as in the May–Holling–Tanner model [4] and other models recently analyzed [28, 45].

In [37] it is affirmed that Leslie models can lead to anomalies in their predictions since they predict that even at very low prey density, when the consumption rate by an individual predator is essentially zero, predator population can nevertheless increase if predator population size is even smaller than prey population size. However, Leslie models are recently employed to model vole-weasel dynamics with a parameter time dependent [25], and the autonomous model proposed in [24] is analyzed in [45]. This scheme of modeling differs from a more common Gause-type model [40] in which the predator equation is based on the mass action principle, since the numerical response is dependent on functional response.

Although it may seem that the two aspects considered in the model contradict each other since the prey population exhibits the Allee effect for low densities, while a nonmonotone functional response is suggested for the aggregation (group defense) when the prey population size is large, it is known that predation induces an Allee effect.

Strikingly, a wide range of predator-driven Allee effects have been reported (see Table 2 in [22]); in particular, there is the case of the Atlantic cod (*Gadus morhua*) that forms schools during the day, since commercial fishing (man as predator) provokes stock collapse because a higher proportion of this aggregative population is caught per unit effort when population declines [10].

Also, for obligately cooperative breeders as the African wild dog (*Lycaon pictus*) and meerkat (*Suricata suricatta*), there is a similar situation, because juvenile survival is lower in small groups than large groups in areas with high predator densities but lower in large groups than small groups in areas with low predator densities [14, 22].

From a mathematical point of view, simple models for the Allee effect may reveal a lot about its dynamics, and, reciprocally, different nonequivalent dynamics for the same model will have different biological interpretations. For instance, in our model we obtain the existence of a subset of parameter values for which three limit cycles appear but surrounding different equilibrium points. These limit cycles are not only periodic solutions of the system but also along the attractive behavior of the origin, allowing the phenomenon of *multistability* of our predator-prey system, that is, the existence of four ω-limit sets in the first quadrant.

This paper is structured as follows: The model and the main results are presented in section 2. In section 3 we give the proofs of the main statements. Finally, an interpretation of the results is given in section 4, complemented with some numerical simulations.

2. The model. Let us consider the bidimensional system of ordinary differential equations

$$X_\mu : \begin{cases} \dot{x} = \left[r\left(1-\frac{x}{k}\right) - \frac{m}{x+b}\right]x - \frac{qxy}{x^2+a}, \\ \dot{y} = sy\left(1-\frac{y}{nx}\right), \end{cases} \tag{3}$$

where $(x, y) \in \mathcal{A} = \{(x, y) |\, x > 0, y \geq 0\}$ and $\mu = (r, a, b, k, m, n, q, s) \in \mathbb{R}^8_+$. In system (3), $x(t)$ and $y(t)$ denote prey and predator densities, respectively, as functions of time t. Furthermore, the parameters have the following meanings:

(a) r and s are the intrinsic growth rates or biotic potential of the prey and predators, respectively.

(b) q is the maximal predator per capita consumption rate, i.e., the maximum number of prey that can be eaten by a predator in each time unit.
(c) a is the number of prey necessary to achieve one-half of the maximum rate q.
(d) n is a measure of the food quality that the prey provides for conversion into predator births.
(e) k is the prey environment carrying capacity [12].
(f) m and b are constants that indicate the severity of the Allee effect that has been modeled.

In order to describe the dynamics of X_μ we will consider a C^∞-equivalent polynomial extension by the following change of coordinates:

$$\zeta : \mathbb{R}^2 \times \mathbb{R}_0^+ \longrightarrow \mathbb{R}^2 \times \mathbb{R}_0^+ \text{ such that } \zeta(u, v, \tau) = \left(u, nv, \frac{u(u+b)(u^2+a)\,\tau}{s}\right) = (x, y, t), \tag{4}$$

where $\det D\zeta(u,v,\tau) = \frac{nu(u+b)(u^2+a)}{s} > 0$.

Also, let us consider new parameters given by $\varphi_1 : \mathbb{R}_+^8 \longrightarrow \mathbb{R} \times \mathbb{R}_+^7$,

$$\varphi_1(r, a, b, k, m, n, q, s) = \left(\frac{br-m}{s}, a, b, k, \frac{m}{s}, n, \frac{nq}{s}, s\right) := (L, a, b, k, M, n, c, s),$$

with Jacobian $\det D\varphi_1(r, a, b, k, m, n, q, s) = \frac{s^3}{bn} > 0$, which says that φ_1 is invertible. For simplicity, let us rename the new coordinates $u \to x, v \to y$. Then in the new coordinates we have the vector field $Y_\eta = \zeta_* X_\mu = (D\zeta)^{-1} \circ X_\mu \circ \zeta$, where

$$Y_\eta : \begin{cases} \dot{x} = x\left((a+x^2)\left[\frac{(b+x)(L+M)}{b}\left(1-\frac{x}{k}\right) - M\right]x - cxy(b+x)\right), \\ \dot{y} = (b+x)(a+x^2)(x-y)y, \end{cases} \tag{5}$$

with $(x, y) \in \bar{\Omega} = \{(x,y) | x, y \geq 0\}$ and the new vector of parameters $\eta = (L, a, b, k, M, c) \in \mathcal{D}_0$ is given by the natural projection, where

$$\mathcal{D}_0 = \left\{\eta \in \mathbb{R} \times \mathbb{R}_+^5 \,|\, M + L > 0\right\}. \tag{6}$$

Notice that vector field (5) is a polynomial extension of the original system (3) to the whole first quadrant $\overline{\Omega}$ including axis $x = 0$. Moreover, Y_η and X_μ are clearly C^∞-equivalent in Ω.

First, we study the local behavior of singularities in the coordinate axis, that is, in absence of prey and predator, respectively. In system (5), it is immediate that $Y_\eta(0, y) = -aby^2 \frac{\partial}{\partial y}$. Then the $x = 0$ axis in $\overline{\Omega}$ is an invariant manifold of vector field (5) and the origin is its only singularity, being an attracting one. Moreover, if we denote $Y_\eta(x, y) = P(x, y)\frac{\partial}{\partial x} + Q(x, y)\frac{\partial}{\partial y}$, we have $Y_\eta(x, 0) = P(x, 0)\frac{\partial}{\partial x}$. Hence, the $y = 0$ axis in $\overline{\Omega}$ is also an invariant manifold of (5).

For vector field Y_η let us define $Sing(E)$ and $Sing(\Omega)$ as the sets of singularities of (5) in the $y = 0$ axis in $\overline{\Omega}$ and in the open set Ω, respectively. In order to describe the dynamics of singularities in $Sing(E)$, we consider the following regions in parameter space $\mathcal{D}_0$:

$\Lambda_1 = \Delta^{-1}(-\infty, 0) \cup \left(\Delta^{-1}(0, \infty) \cap \Theta^{-1}(0, \infty) \cap \Upsilon^{-1}(-\infty, 0)\right)$;	
$\Lambda_2 = \Delta^{-1}(0, \infty) \cap \Theta^{-1}(\infty, 0) \cap \Upsilon^{-1}(-\infty, 0)$;	$\Lambda_3 = \{\eta \in \mathcal{D}_0 \,\vert\, 0 < L < b\}$;
$\Lambda_4 = \{\eta \in \mathcal{D}_0 \,\vert\, 0 < b \leq L\}$;	$D = \Delta^{-1}(0) \cap \Theta^{-1}(0, \infty)$;
$\Theta_+ = \Theta^{-1}[0, \infty) \cap \Upsilon^{-1}(0)$;	$\Theta_- = \Theta^{-1}(-\infty, 0) \cap \Upsilon^{-1}(0)$;
$\Gamma_+ = \Gamma^{-1}(0, \infty) \cap \Upsilon^{-1}(0)$;	$\Gamma_- = \Gamma^{-1}(-\infty, 0] \cap \Upsilon^{-1}(0)$;

(7)

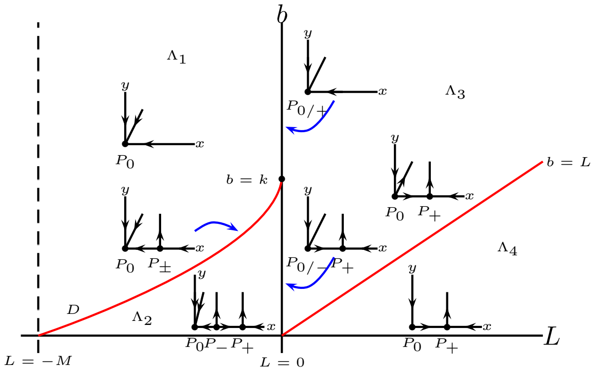


FIG. 1. *Bifurcation diagram in plane $Lb = \{\eta \in \mathcal{D}_0 |\, a, k, M, c \text{ are constant}\}$ in parameter space for singularities in the $y = 0$ axis.*

where $\Upsilon(\eta) = sign(L)$, $\Theta(\eta) = b - k$, and

$$\Delta(\eta) = (M + L)(k - b)^2 + 4bkL,$$
$$\Gamma(\eta) = -2kcb + a\sqrt{M}\left(-a\sqrt{M} + \sqrt{aM + 4k^2c}\right).$$

Moreover, in the xy plane let us consider the points $P_0 = (0, 0)$, $P_+ = (x_+, 0)$, and $P_- = (x_-, 0)$, with

$$x_\pm = \frac{(M + L)(k - b) \pm \sqrt{(M + L)\Delta(\eta)}}{2(M + L)}. \tag{8}$$

LEMMA 1. *For vector field Y_η the origin P_0 is a non-hyperbolic singularity and it satisfies the following:*

(i) *It has two hyperbolic sectors divided by a repulsing separatrix if $\eta \in \Gamma_+ \cap \Theta_+$.*
(ii) *It has only a hyperbolic sector if $\eta \in \Lambda_4$.*
(iii) *It has a hyperbolic sector and a repulsing parabolic sector if $\eta \in \Lambda_3 \cup (\Theta_- \cap \Gamma_+)$.*
(iv) *It has a hyperbolic sector and an attracting parabolic sector if $\eta \in \Theta_- \cap \Gamma_-$.*
(v) *It is a local attractor if $\eta \in D \cup \Lambda_1 \cup \Lambda_2 \cup (\Theta_+ \cap \Lambda_-)$.*

A qualitative diagram of these results in plane $Lb = \{\eta \in \mathcal{D}_0 |\, a, k, M, c \text{ are constant}\}$ is shown in Figure 1.

LEMMA 2. *For vector field Y_η the following statements hold:*

(i) *$P_+ \in Sing(E)$ if $\eta \in D \cup \Theta_- \cup \Lambda_3 \cup \Lambda_4$. Moreover, P_+ is a saddle node if $\eta \in D$; otherwise, it is a hyperbolic saddle.*
(ii) *$P_- \in Sing(E)$ if $\eta \in \Lambda_2$. Moreover, P_- is a hyperbolic repulsing node.*

A qualitative diagram of these results in plane $Lb = \{\eta \in \mathcal{D}_0 |\, a, k, M, c \text{ are constant}\}$ is shown in Figure 1.

About the boundness of the system, in [1] it is proven that no trajectory of vector field Y_η has the infinity as ω-limit. Since the coordinate axes are invariant, then our system is bounded. We now state the same result but with a different approach.

THEOREM 3. *Consider $\mathcal{A}_w = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq w, 0 \leq y\}$. For every $\eta \in \mathcal{D}_0$, there is a $w^* \geq 0$ such that if $w > w^*$, then $\mathcal{A}_w$ is a trapping domain for system (5), meaning that it is invariant for positive time evolution and also captures all trajectories starting in $\overline{\Omega}$.*

We take care now of the dynamics at the interior of the first quadrant. It is clear that every singularity of system (5) in the open quadrant Ω must be over the diagonal $y = x$. Furthermore, over the isocline $y = x$ the vector field Y_η has the form $Y_\eta(x, x) = x^2 p_4(x)\frac{\partial}{\partial x}$, where $p_4(x)$ is a polynomial of degree 4 in x. Therefore, the existence of equilibrium points in the open first quadrant is given by the existence of real and positive roots of the polynomial $p_4(x)$. In order to describe the bifurcation diagram for an equilibrium point of Y_η in Ω, we change parameters $\eta = (L, a, b, k, M, c)$ to get a new vector $\xi = (L, a, b, k, V, q)$ by means of the transformation

$$\varphi_2 : \mathbb{R}^6 \longrightarrow \mathbb{R}^6, (L, a, b, k, V, q) \mapsto (L, a, b, k, M(\xi), c(\xi)), \tag{9}$$

with

$$\boxed{\begin{aligned} c(\xi) &= \frac{aLq^4(b-k+2q)-q^6(2b^2+kL-b(2k+L-4q)-2(k+L)q+2q^2)}{kq^6} - \frac{(b-k+q)V}{kq^6b^2(b+q)^5(a+q^2)^2(a+3q^2)},\\ M(\xi) &= \frac{-bq^4(b+q)^6(a+q^2)^2(a+3q^2)(aL+(L-2b)q^2)+V}{bq^6(b+q)^4(a+q^2)^3(a+3q^2)}. \end{aligned}}$$

Let $\mathcal{D}_1 = \varphi_2^{-1}(\mathcal{D}_0)$ be the new admissible region in parameter space, and let us consider the projection

$$\pi : \mathbb{R}^6 \longrightarrow \mathbb{R}^3 \quad \text{such that} \quad \pi(L, a, b, k, q, V) = (L, b, k).$$

If we name

$$\mathcal{D}_2 = \pi\left(\mathcal{D}_1 \cap \{(L, a, b, k, q, V)|\ a = 3b^2, q = b, V > 0\}\right) \cap E^{-1}(0, \infty) \subset \mathbb{R} \times \mathbb{R}^2_+,$$

with

$$E(L, b, k) = -18432b^{19}L + V,$$

it is easy to check that $\mathcal{D}_2 = F^{-1}(0, \infty) \cap E^{-1}(0, \infty)$, with

$$F(L, b, k) = 12288b^{19}(b - 2L) + V.$$

Changing the time $t \mapsto bkt$, let us call Y_ξ to the new vector field qualitatively equivalent to (5). Now, it is straightforward to check that there is a singularity in the isocline $y = x$ at the point with coordinates (b, b). In order to describe the bifurcation diagram of this singularity of vector field Y_ξ in parameter space (L, b, k), we consider the projection (see Figure 2)

$$\pi_R : S^2_R \longrightarrow \mathbb{R}^2 \quad \text{such that} \quad \pi_R(L, b, k) = (L, b),$$

where $S^2_R = \{(L, b, k)|\, L^2 + b^2 + k^2 = R^2, b, k, R > 0\}$. Since function $F(L, b, k)$ does not depend on k, it is clear that in $\mathcal{D}_2 \subset \mathbb{R} \times \mathbb{R}^2_+$ the surface $F^{-1}(0)$ is a straight half-cylinder with axis parallel to the k axis, so $\pi_R\left(F^{-1}(0) \cap S^2_R\right)$ is exactly the curve

$$\mathcal{F} = \{(L, b)|\, F(L, b, k) = 0,\ k > 0\} \cap \pi_R\left(\mathcal{D}_2 \cap S^2_R\right),$$

which is located in the interior of the first quadrant of plane Lb because equation $F(L, b, k) = 0$ defines implicitly $L = \frac{b}{2} + \frac{V}{24576b^{19}} > 0$. Furthermore, $(0, 0) \in F^{-1}(0, \infty)$, so the origin is in the admissible region for parameters in the upper half-plane Lb. The same arguments follow to sketch the curve $\pi_R\left(E^{-1}(0) \cap S^2_R\right)$. With this, the qualitative shape of domain $\pi_R(\mathcal{D}_2 \cap S^2_R)$ is shown in Figure 3(a).

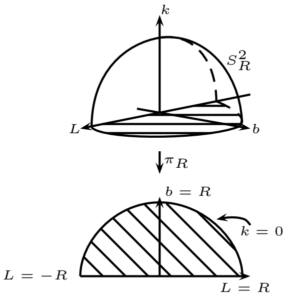
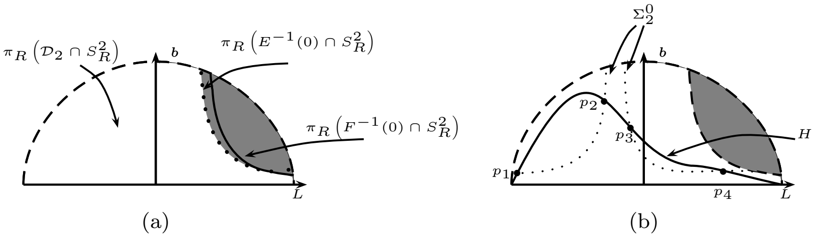


FIG. 2. *Projection π_R from the sphere S^2_R to the plane Lb.*


$\pi_R\left(\mathcal{D}_2 \cap S^2_R\right)$
b
$\pi_R\left(E^{-1}(0) \cap S^2_R\right)$
$\pi_R\left(F^{-1}(0) \cap S^2_R\right)$
L
Σ^0_2
b
p_2
p_3
H
p_1
p_4
L

(a) (b)

FIG. 3. (a) *Qualitative shape of parameters domain* $\pi_R\left(\mathcal{D}_2 \cap S^2_R\right)$. (b) *Qualitative shape of curves H and $\Sigma^2_0(R)$ in parameters domain.*

Finally, in $\pi_R\left(\mathcal{D}_2 \cap S^2_R\right)$ in plane Lb, we define the sets

$$\begin{aligned} \Sigma_1^{sg(\epsilon)}(R) &= \pi_R\left(\{(L,b,k) \in \mathcal{D}_2 |\, sg(T(L,b,K)) = sg(\epsilon)\} \cap S^2_R\right), \\ \Sigma_2^{sg(\epsilon)}(R) &= \pi_R\left(\{(L,b,k) \in \mathcal{D}_2 |\, sg(G(L,b,K)) = sg(\epsilon)\} \cap S^2_R\right), \\ H &= \Sigma_1^0(R), \end{aligned} \tag{10}$$

with

$$\begin{aligned} T(L,b,k) &= -36864b^{21} + 55296b^{20}L - 18432b^{19}kL - 3bV + kV, \\ G(L,b,k) &= 1811939328b^{39}(b+L)(b+2L) - 8bV^2, \end{aligned}$$

and the qualitative shape of curves H and $\Sigma^0_2(R)$, in terms of R, is as in Figure 3(b). Moreover, $H \cap \Sigma^0_2(R)$ consists of four points as in Figure 3(b), that is,

$$H \cap \Sigma^0_2(R) = \{p_i = (L_i, b_i); i = 1, \ldots, 4\}. \tag{11}$$

LEMMA 4. *The singularity (b,b) of vector field Y_ξ is an order two weak focus if $(L,b) \in H \cap \Sigma^0_2(R) = \{p_i = (L_i, b_i); i = 1, \ldots, 4\}$. Moreover, the focus is repellor at p_1 and p_2 and attractor at p_3 and p_4.*

THEOREM 5.

(i) *The system Y_ξ has a bifurcation diagram as indicated in Figures* 4(a)–(d) *in a neighborhood of $H \cap \Sigma^0_2(R)$.*

(ii) *The bifurcation curve S_{34} of semistable limit cycle of the system Y_ξ in the parameter space $\Sigma^-_1(R)$ is as indicated in Figure* 4(d).

(iii) *The bifurcation curve HC of a homoclinic loop surrounding the focus (b,b) of system Y_ξ is as indicated in Figure* 4(a), *in a tubular neighborhood of curve H in sector $\Sigma^-_2(R)$ in parameter space.*

LEMMA 6. *In parameter space $\xi = (L,a,b,k,q,V) \in \mathcal{D}_1$ of system Y_ξ, there exists an open subset $\Upsilon \subset \left(\pi^{-1} \circ \pi_R^{-1}(p_3)\right) \cap \{\xi \,|\, L < 0\}$, where the following statements hold:*

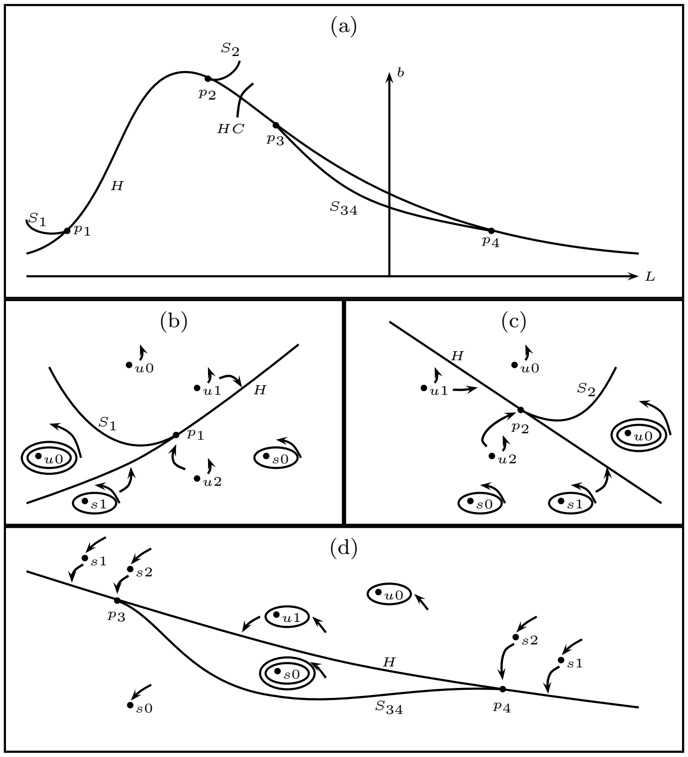


FIG. 4. (a) *Bifurcation curves for focus* (b, b) *of vector field* Y_ξ*, where* H *is a Hopf bifurcation curve,* S_1, S_2*, and* S_{34} *are semistable limit cycle bifurcation curves, and* HC *is a Homoclinic bifurcation curve.* (b) *Bifurcation diagram for focus* (b, b) *in a neighborhood of* $p_1 \in H$*.* (c) *Bifurcation diagram for focus* (b, b) *in a neighborhood of* $p_2 \in H$*.* (d) *Bifurcation diagram for focus* (b, b) *in a tubular neighborhood of curve* H *between* p_3 *and* p_4*. Notation* $\bullet_{ij}$ *means a focus with local stability* $i \in \{s, u\}$ *and weakness* $j \in \{0, 1, 2\}$*.*

(i) Y_ξ *has two foci* $(b, b), (x_f, x_f)$ *and two saddle points* $(x_s, x_s), (x_N, x_N)$ *in the interior of the first quadrant* Ω*, where* $0 < x_s < b < x_N < x_f$*.*

(ii) *There exists an open subset* $\mathcal{O} \in \mathbb{R}_+$ *and a continuous function* $\mathfrak{H} : \mathcal{O} \longrightarrow \mathbb{R}_+$ *such that if* $V = \mathfrak{H}(b)$*, then there is a homoclinic loop* $\mathcal{L}_0$ *for vector field* Y_ξ *passing through saddle* (x_N, x_N) *and surrounding focus* (x_f, x_f)*, with* (x_N, x_N) *and* (x_f, x_f) *as in part* (i)*. Moreover, this homoclinic orbit is inner unstable.*

LEMMA 7. *Let* $\Upsilon, \mathfrak{H}$*, and* $\mathcal{L}_0$ *be as in Lemma* 6*. There exist* $\delta > 0$ *and* $\varepsilon > 0$ *such that, for every* b, V *with* $|\mathfrak{H}(b) - V| < \varepsilon$*, the vector field* Y_ξ *has, at most, one limit cycle* $\mathcal{L}$ *in the inner* δ*-neighborhood of* $\mathcal{L}_0$*, and* $\mathcal{L}$ *is unstable.*

THEOREM 8. *There exists an open subset in parameter space such that vector field* Y_η *in* (5) *has three hyperbolic limit cycles in the interior of the first quadrant* Ω*. Two of them are infinitesimal cycles surrounding a hyperbolic attracting focus* (b, b)*; moreover, the outermost limit cycle is stable and the innermost unstable. The third limit cycle is unstable and surrounds a hyperbolic attracting focus* (x_f, x_f)*, with* $0 < b < x_f$*.*

3. Proofs of the main results. The proofs of Lemmas 1 and 2 are straightforward and follow from the blowing-up method [17] for $(0, 0)$ and the central manifold and Hartman's theorems for P_+ and P_-; for details, see [1].

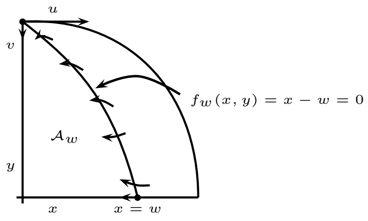


FIG. 5. *The set* $\mathcal{A}_w = f_w^{-1}(-\infty, 0) \cap \overline{\Omega}$ *is an invariant set under* $\Phi_{Y_\eta}(t; \cdot)$, *as stated in the proof of Theorem* 3.

Proof of Theorem 3. We will prove that for every $\eta \in \mathcal{D}_0$ (see (6) for definition), there exists some $w = w(\eta) > 0$ such that the integral curves $\Phi_{Y_\eta}(t; \cdot)$ of vector field Y_η, for $t > 0$, enter an invariant subset $\mathcal{A}_w \subset \overline{\Omega}$ and do not leave again.

Let us define the function $f_w : \mathbb{R}^2 \longrightarrow \mathbb{R}$ given by $f_w(x,y) = x - w$. For each $w > 0$ consider the line

$$L_w = f_w^{-1}(0) \cap \overline{\Omega}.$$

If we consider the Poincaré compactification of system (5) given by

$$\Psi_y : \mathbb{R}^2 \times \mathbb{R}_0^+ \longrightarrow \mathbb{R}^2 \times \mathbb{R}_0^+, \;\; \Psi_y(u, v, \tau) = (u/v, 1/v, \tau) = (x, y, t),$$

it is straightforward to check that the line $\Psi_y^{-1}(L_w)$ contains the point $(u,v) = (0,0)$. Hence, we will show that for w sufficiently large, in the Poincaré compactification of system (5), the compact set $\mathcal{A}_w = f_w^{-1}(-\infty, 0) \cap \overline{\Omega}$ is an invariant set under $\Phi_{Y_\eta}(t; \cdot)$ (see Figure 5).

Recall that the axes $x = 0$ and $y = 0$ are invariant under $\Phi_{Y_\eta}(t; \cdot)$, so it is sufficient to prove that the scalar product $Y_\eta \cdot \nabla f_w < 0$ in L_w for appropriate $w > 0$ as in Figure 5. We have that for $(x,y) \in L_w$

$$Y_\eta \cdot \nabla f_w \,|_{x=w} = -cw^2(b+w)y + \frac{w^2\left(a+w^2\right)}{bk}\,\phi(w),$$

where $\phi(w) = -(M+L)w^2 + (k-b)(M+L)w + bkL$. The discriminant of ϕ as a quadratic polynomial in w is $(M+L)\Delta(\eta)$ (see (7) for definition). Therefore, since $M + L > 0$ in $\mathcal{D}_0$, if $\Delta < 0$, then $Y_\eta \cdot \nabla f_u \,|_{x=w} < 0$ for every $w > 0$. On the other side, if $\Delta \geq 0$, the roots of $\phi(w)$ are x_- and x_+ (see (8) for definition); hence, the statement holds for every $w > x_+$ if $x_+ > 0$ and for every $w > 0$ if $x_+ \leq 0$. This concludes the proof of Theorem 3. □

Proof of Lemma 4. Let $(L,b) \in \pi_R(\mathcal{D}_2 \cap S_R^2)$. In order to study the singularity (b,b) in vector field Y_ξ, we translate (b,b) to the origin and change the time by means of

$$\zeta : \mathbb{R}^2 \times \mathbb{R}_0^+ \longrightarrow \mathbb{R}^2 \times \mathbb{R}_0^+,$$
$$(x,y,t) \mapsto \left(x+b, y+b, \frac{8b^3E(L,b,k)}{768b^{15}}\,t\right),$$

where

$$E(L,b,k) = -18432b^{19}L + V.$$

So if $(L,b,k) \in E^{-1}(0,\infty)$, ζ is a $\mathcal{C}^\infty$-equivalence. Let $Z_\xi = \zeta_* Y_\xi$; then we have

$$\mathrm{tr} DZ_\xi(0,0) = \frac{8E(L,b,k)T(L,b,k)}{589824 b^{25}},$$

with

$$T(L,b,k) = -\left(36864b^{21} - 55296b^{20}L + 18432b^{19}kL + 3bV - kV\right).$$

Moreover, if $(L,b) \in H = \pi_R(T^{-1}(0) \cap S_R^2)$, we have $\det DZ_\xi(0,0) = V > 0$, so the origin is a weak focus of order, at least, one for Z_ξ. In order to recognize the topological type of the origin as $(L,b) \in H$, we bring Z_ξ onto its Jordan canonical form at $(0,0)$. For that, consider the $\mathcal{C}^\infty$-equivalence $\varphi : \mathbb{R}^2 \times \mathbb{R}_0^+ \longrightarrow \mathbb{R}^2 \times \mathbb{R}_0^+$,

$$\varphi(u,v,\tau) = \left(192b^{10}u - \sqrt{V}\,v, 192b^{10}u, \tau/\sqrt{V}\right) = (x,y,t).$$

The qualitatively equivalent vector field in the new coordinates is $Z_\xi^J = \varphi_* Z_\xi$, and we have in a neighborhood of the origin

$$Z_\xi^J(u,v) = \left(-v + \sum_{i,j=2}^{5} A_{i,j}u^i v^j + \text{H.O.T.}\right)\frac{\partial}{\partial u} + \left(u + \sum_{i,j=2}^{5} B_{i,j}u^i v^j + \text{H.O.T.}\right)\frac{\partial}{\partial v},$$

where H.O.T. denotes the higher order term and $A_{i,j} = A_{i,j}(b,V,L)$, $B_{i,j} = B_{i,j}(b,V,L)$.

For $j = 0,1,2$, let l_j be the first three Lyapunov quantities at the origin [2, 7, 33] of the vector field Z_ξ^J. Since the trace of the linear part of the vector field at the origin is zero, we have that $l_0 = 0$. Now, as l_1 depends on the 3-jet of Z_ξ^J, see [17], then

$$l_1 = (A_{02}A_{11} + A_{12} + A_{11}A_{20} + 3A_{30} + 2A_{02}B_{02} + 3B_{03} - B_{02}B_{11} - 2A_{20}B_{20} - B_{11}B_{20} + B_{21})/8.$$

Using the Mathematica software [41] we have

$$l_1 = \frac{36864b^{20} + V}{8192b^{13}V^{3/2}}\, G(L,b,k),$$

with

$$G(L,b,k) = 1811939328b^{39}(b+L)(b+2L) - 8bV^2.$$

Thereby, if $(L,b) \in H \cap \Sigma_2^0(R)$ (see (10) and (11)), the weakness of the origin depends only on l_2. On the other hand, it is known that l_2 depends on the 5-jet of Z_ξ^J. Then, again using the Mathematica software, as $l_0 = l_1 = 0$, we obtain

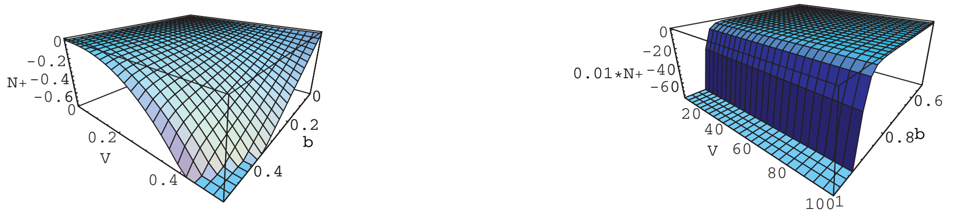
$$l_{2_\pm} = \frac{36864b^{20} + V}{98304b^{15}V^{3/2}} N_\pm,$$

where

$$\begin{aligned} N_\pm = -\big(&-221259535220736b^{61} - 135895449606b^{41}V - 1963008b^{21}V^2 - 57bV^3 \\ &\pm\sqrt{3}\left(303499837744b^{40} + 2211840b^{20}V + 43V^2\right)\sqrt{28311552b^{42} + b^2V^2}\,\big) \end{aligned}$$

and the sign ± denotes the branch

$$L_\pm = \frac{-27648b^{39} \pm \sqrt{3}\sqrt{g(b)}}{36864b^{38}} \tag{12}$$



(a) (b)

FIG. 6. *Graphics of the function* $N_+ = N_+(b, V)$ *where we can see that* $N_+ < 0$*:* (a) *graphic of* N_+ *for* $(b, V) \in [0, 0.5] \times [0, 0.5]$*;* (b) *graphic of* $10^{-2} \times N_+$ *for* $(b, V) \in [0.6, 1] \times [10, 100]$.

of curve $\Sigma_2^0(R)$ given implicitly by equation $G(L, b, k) = 0$, where $g(b) = 28311552b^{78} + b^{38}V^2 > 0$. Since the branch L_- intersects curve H at points p_1 and p_2 (see Figure 3(b)), then if $(L, b) \in \{p_1, p_2\} \subset H \cap \Sigma_2^0(R)$, $l_2 > 0$ and the origin is a repellor weak focus of order two for Z_ξ^J.

On the other hand, the branch L_+ intersects curve H at points p_3 and p_4 (see Figure 3(b)), and by looking at the graphic of N_+ as a function of b and V it can be verified that $N_+ < 0$. Figure 6(a) shows the graphic of N_+ for the range $(b, V) \in [0, 0.5] \times [0, 0.5]$, meanwhile in Figure 6(b) there is the graphic of $10^{-2} \times N_+$ for $(b, V) \in [0.6, 1] \times [10, 100]$, made with the software Mathematica [41]; here it can be seen that if $b > 0, V > 0$, as in our case, then by continuity $N_+ < 0$. Therefore, if $(L, b) \in \{p_3, p_4\} \subset H \cap \Sigma_2^0(R)$, $l_2 < 0$ and the origin is an attracting weak focus of order two for Z_ξ^J. □

Proof of Theorem 5(i). It is clear that $T^{-1}(0)$ and S_R^2 are transversal manifolds in the parameter space and, by Lemma 4, there are four points

$$p_1, p_2, p_3, p_4 \in \pi_R \left(T^{-1}(0) \cap S_R^2\right) = H,$$

where H is a Hopf bifurcation curve and where the singularity (b, b) is a repellor weak focus of order 2 of system Y_ξ at p_1 and p_2 and an attractor weak focus of order two at p_3 and p_4; i.e., the eigenvalues of $DY_\xi(b, b)$ are on the imaginary axes and nonzero. Moreover, $p_1, p_2, p_3,$ and p_4 have codimension two, since the Lyapunov quantities $l_0 = l_1 = 0$ and $l_2 \neq 0$.

Thereby, let $(L, b) \in H - \{p_1, p_2, p_3, p_4\}$. Then we have the following (see Figure 7):

(a) In a neighborhood of p_4, when (L, b) is below $l_1^{-1}(0)$, the singularity (b, b) of system Y_ξ is an order one attractor weak focus, because $l_1 < 0$. Therefore, if $\epsilon > 0$ is sufficiently small, the point $(L + \epsilon, b + \epsilon)$ is over the curve H; then the stability of system Y_ξ at the singularity (b, b) is reversed; hence, a unique hyperbolic stable limit cycle bifurcates (Hopf bifurcation).

(b) When (L, b) is above $l_1^{-1}(0)$ in a neighborhood of p_4, the singularity (b, b) is an order one repellor weak focus surrounded by a hyperbolic stable limit cycle, because $l_1 > 0$; then Hopf bifurcation from p_4 occurs. Therefore, if $\epsilon > 0$ is sufficiently small, the point $(L + \epsilon, b)$ is over the curve H; hence, the stability of system Y_ξ at the singularity (b, b) does not change and the limit cycle persists because it is hyperbolic.

(c) When the point (L, b) is as in case (b), for $\epsilon > 0$ sufficiently small, the point $(L - \epsilon, b)$ is under the curve H; then the stability of system Y_ξ at the singularity (b, b) is reversed; hence, a new Hopf bifurcation occurs and a

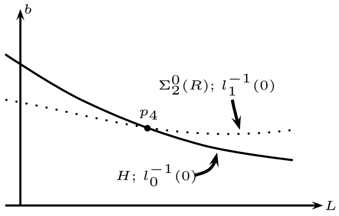


FIG. 7. *Hopf bifurcation curve H near point p_4 in parameter space, where the singularity (b, b) of vector field Y_ξ is an order two attracting weak focus.*

second hyperbolic limit cycle bifurcates, which is surrounded by a hyperbolic stable limit cycle.

By normal forms theory [17, 23], p_4 has a 2-parameter versal unfolding. Under the same hypothesis that we have here, Takens [35] and Arrowsmith and Place [3] describe in detail the bifurcation diagram for codimension two singularity type. By using their results, we have that there exists a neighborhood $\mathcal{V}_\delta(p_4)$, with $\delta > 0$, such that it has a diagram as in Figure 4, where S_{34} is a curve in which the unstable and stable limit cycles collapse (semistable limit cycle). The analysis near points p_1, p_2, and p_3 follows in the same way.

Proof of Theorem 5(ii). From part (i) of this Theorem, there is an open set $\mathcal{C}(R) \subset \Sigma_1^-(R)$ where system Y_ξ has two hyperbolic limit cycles surrounding the singularity (b, b). Moreover, $\mathcal{C}(R)$ is upperly bounded by the segment of curve H between p_3 and p_4, because at any point in that segment focus (b, b) is surrounded by a unique infinitesimal limit cycle.

On the other hand, at $\Sigma_1^-(R) \backslash \mathcal{C}(R)$, the singularity (b, b) has no infinitesimal limit cycle in its vicinity, so $\mathcal{C}(R)$ is bounded and the lower bound must correspond to a bifurcation curve where both limit cycles collapse into a semistable limit cycle; hence, by versality of the unfolding of the singularity (b, b) at p_3 and p_4 described in part (i) (see [35, 3]), it is clear that the semistable limit cycle bifurcation curve S_{34} must be as in Figure 4(d).

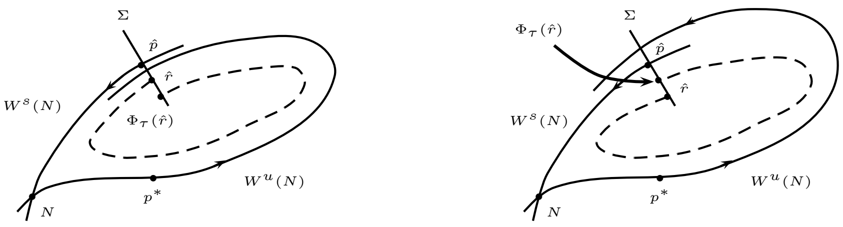
Proof of Theorem 5(iii). By Hartman–Grobman's theorem it is straightforward to see that vector field Y_ξ has a hyperbolic saddle singularity $P_s = (s, s)$ with $0 < s < b$ if $(L, b) \in H \cap \Sigma_2^-(R)$.

Let $(L, b) \in H \cap \Sigma_2^-(R)$ in a neighborhood of p_3. By part (i), (b, b) is an attracting weak focus of order one with no limit cycle in its vicinity, and it can be checked that $P_s \in \alpha-\lim(b, b)$. As we move along segment $H \cap \Sigma_2^-(R)$ towards p_2, the topological type of focus (b, b) remains the same according part (i); however, if $(L, b) \in H \cap \Sigma_2^-(R)$ in a neighborhood of p_2, $P_s \in (\alpha-\lim(b, b))^c$ and an infinitesimal limit cycle exists surrounding (b, b). Therefore, there exist $\delta > 0$ sufficiently small and a differentiable function

$$\Psi : V_\delta(H) \longrightarrow \mathbb{R} \quad \text{such that} \quad (L, b) \mapsto \Psi(L, b),$$

defined in a tubular neighborhood $V_\delta(H)$ of radius δ of segment $H \cap \Sigma_2^-(R)$ such that vector field Y_ξ has a homoclinic loop passing through the saddle P_s and surrounding focus (b, b), if $(L, b) \in HC = \Psi^{-1}(0)$.

Remark. An extended and deeper proof of the existence of a homoclinic loop surrounding another singularity is given in Lemma 6. Nevertheless, the same arguments might have been used in this case as well. □



(a) (b)

FIG. 8. *Poincaré map for orbits near* $W^s(N)$ *in two cases:* (a) $W^u(N) \cap W^s(x_f) \neq \emptyset$; (b) $W^u(N) \cap W^s(x_f) = \emptyset$.

Proof of Lemma 6(i). Let $(L, b) = p_3 \in H \cap \Sigma_2^0$. It is easy to check that Y_ξ has four singularities in isocline $y = x$ at Ω, if $L < 0$. According to Lemma 4, one of these equilibria is an attracting weak focus of order two in (b, b). Due to Hartman–Grobman's theorem, it is straightforward to see that the other singularities are a hyperbolic attracting focus in (x_f, x_f) and two saddles (x_s, x_s), (x_N, x_N) with $0 < x_s < b < x_N < x_f$. The statement follows directly.

Proof of Lemma 6(ii). Let (x_N, x_N) and (x_f, x_f) be as in the proof of part (i). If we name $N = (x_N, x_N)$, let us consider the function

$$\mathfrak{D}(N) = \nabla \cdot Y_\xi(N),$$

and let $\Phi_t = \Phi_{Y_\xi}(t; \cdot)$ be the flow of Y_ξ and $p^* = (x^*, y^*)$ be a point in $W^u(N)$ with $x_N < x^*$. Let Σ be a one-dimensional local cross section to $W^s(N) \cup W^u(N)$ as in Figure 8. Let $\hat{p} = (\hat{p}_1, \hat{p}_2)$ the unique point of $\Sigma \cap W^s(N)$ which never returns to Σ by the flow Φ_t with $0 < t < \infty$, and let $U \subset \Sigma$ be some neighborhood of $\hat{p}$. In Σ let us choose local coordinates given by the chart $h : \Sigma \longrightarrow \mathbb{R}$, $h(x, y) = y - \hat{p}_2$.

Let us consider the *Poincaré map* $\mathcal{P} : U \longrightarrow \Sigma$ defined for a point $r \in U$ by $\mathcal{P}(r) = \Phi_\tau(r)$, where $\tau = \tau(r)$ is the time taken for the orbit $\Phi_\tau(r)$ based at r to first return to Σ. Now consider the displacement function $d : U \longrightarrow \mathbb{R}$ given by

$$r \mapsto d(r) = \mathcal{P}(r) - r.$$

Finally, let $\hat{r}$ be a point in U located below $\hat{p}$ as in Figure 8.

Let $b = 1$ and $V = 300$. Then it can be seen by means of numerical simulations on Matlab [30] that $\mathfrak{D}(N) > 0, W^u(N) \cap W^s(x_f) \neq \emptyset, W^u(N) \cap W^s(N) = \emptyset$, and $\omega - \lim \Phi_t(p^*) = \{(x_f, x_f)\}$. In the local coordinates given by h this implies that

$$d(\hat{r}) < 0. \tag{13}$$

A qualitatively equivalent situation is shown in Figure 8(a).

Instead, if $b = 1$ and $V = 400$, it can be checked (numerically as well) that $\mathfrak{D}(N) > 0$, $W^u(N) \cap W^s(x_f) = \emptyset$, and $W^u(N) \cap W^s(N) = \emptyset$; there exists $y^\dagger > x_N$ such that the point $(x_N, y^\dagger) \in \Phi_t(p^*) \subset W^u(N)$. Therefore, we have (see Figure 8(b))

$$d(\hat{r}) > 0. \tag{14}$$

Hence, by the continuous dependence of the orbits of the vector field on parameters, there exists V^*, $300 < V^* < 400$, such that for $b = 1$ and $V = V^*$ we have $W^u(N) \cap W^s(N) \neq \emptyset$; that is, there is a homoclinic loop $\mathcal{L}_0 \subset W^u(N) \cap W^s(N)$.

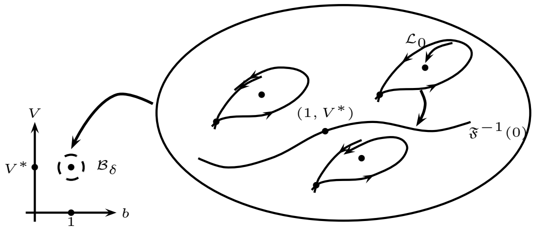


FIG. 9. *Existence of a homoclinic orbit along the curve* $\mathfrak{F}^{-1}(0)$ *in plane* bV.

Therefore, there exists $\delta > 0$ such that in a δ-ball $\mathcal{B}_\delta(1, V^*) \subset \mathbb{R}^2_+$ of the point $(1, V^*)$ in plane bV in parameter space, there is a locally differentiable function

$$\mathfrak{F} : \mathcal{B}_\delta(1, V^*) \longrightarrow \mathbb{R}, \ \ (b, V) \mapsto \mathfrak{F}(b, V),$$

such that if $(b, V) \in \mathfrak{F}^{-1}(0)$, there exists an inner unstable homoclinic loop $\mathcal{L}_0$ for vector field Y_ξ (see Figure 9). The existence of the function $\mathfrak{H} : \mathcal{O} \longrightarrow \mathbb{R}_+$ in the statement follows from the implicit function theorem, due to the transversality shown in passing from (13) to (14). □

Proof of Lemma 7. Let the function $d : U \longrightarrow \mathbb{R}$, the saddle $N = (x_N, x_N)$, and the point $\hat{r}$ be as in the proof of Lemma 6 (see Figure 8). From the previous lemma, a homoclinic bifurcation (see [29]) occurs as $V = \mathfrak{H}(b)$ since, by continuity, $\mathfrak{D}(N) = \nabla \cdot Y_\xi(N) > 0$, N is a *strong* saddle [29], and $\mathcal{L}_0$ is inner unstable. Then, for U sufficiently small, by carrying out a perturbation on parameters, such that $d(\hat{r}) > 0$, the homoclinic cycle breaks out and an unstable limit cycle appears surrounding the attracting focus (x_f, x_f). □

Proof of Theorem 8. From Lemmas 4 and 7, in parameter space $\xi = (L, a, b, k, q, V) \in \mathcal{D}_1$ of system Y_ξ, there exists an open subset $\Upsilon_{\mathcal{L}} \subset \left(\pi^{-1} \circ \pi_R^{-1}(p_3)\right) \cap \{\xi \,|\, L < 0\}$, where system Y_ξ has four singularities in the interior of the first quadrant Ω: two saddles $(x_s, x_s), (x_N, x_N)$, a two-order attracting weak focus (b, b), and a hyperbolic attracting focus (x_f, x_f) surrounded by an unstable limit cycle $\mathcal{L}$, with $0 < x_s < b < x_N < x_f$. Moreover, this limit cycle is hyperbolic, so it persists in an open subset of $\pi^{-1}(\pi_R^{-1}(\mathcal{C}(R)))$ in parameter space, where $\mathcal{C}(R)$ is the bounded sector in plane Lb (see Figure 4(d)) where system Y_ξ also has two infinitesimal limit cycles surrounding focus (b, b), due to the proof of Theorem 5(ii).

The statement of the theorem is immediate after recalling that vector field Y_ξ is induced by a vector field $\mathcal{C}^\infty$-equivalent to Y_η by means of the local difeomorphism φ_2 in (9), since

$$\det D\varphi_2(\xi) = \frac{V}{b^3 q^{12} (b+q)^{10} (a+q^2)^4 (a+3q^2)^3} > 0,$$

where $\xi \in \mathcal{D}_1 \cap \pi^{-1}(T^{-1}(0))$. □

4. Discussion. In this paper, a predator-prey model with the Allee effect [5, 16, 34, 38] and a Holling-type IV [36] functional response was considered, making a qualitative analysis of a bidimensional system of ordinary differential equations of polynomial type, which is topologically equivalent to the original one, with only six parameters. Employing a reparameterization and a time change, we showed that b, k, and $L = \frac{br-m}{s}$ are the most significative parameters of this predator-prey model. We note that this model predicts that, when $L < 0$, the system has three equilibria

P_0, P_-, and P_+ for which the predator density is null; moreover, at low population densities, both predator and prey populations may disappear, since the equilibrium $P_0 = (0, 0)$ is an attractor; furthermore, P_- is a repelling node, and P_+ is a saddle.

Unlike other models, in this work we show that the weak Allee effect does not imply extinction of species necessarily for certain parameter values, which, as far as we know, is a completely new feature for the continuous model that has not been reported in the above works. As $L = 0$, P_- collapses with the origin and a weak Allee effect is obtained if P_+ remains an isolated singularity. In this case, the model predicts two phenomena depending on parameter values. In the first one, the origin has a separatrix dividing a repelling parabolic sector and a hyperbolic sector; see Lemma 1(iii). This means that every solution with initial conditions (x_0, y_0) in the interior of the first quadrant near the origin ultimately tends to go away from the origin allowing both populations to not extinct.

The second case is more usual, see Lemma 1(iv), and has been observed in previous works [19, 32]; that is, the origin has a parabolic attracting sector and a hyperbolic sector divided by an attracting separatrix, representing a critical line which has to be trespassed for populations in order to survive in time and not extinct.

In this work, we also show in Theorem 8 that, for an open subset of the parameter values, three limit cycles can coexist in the open quadrant Ω. This result has been recently observed [26] but just in general cubic polynomial systems. In our case, only two of these limit cycles are concentric though, since infinitesimal ones are generated by Hopf bifurcation, and the third one is generated by a homoclinic bifurcation [23, 29] and surrounds another focus as stated in Lemma 6 and Theorem 8.

Furthermore, the model can achieve the phenomenon of *multistability* by the existence of four ω-limit sets in the first quadrant as $L < 0$ and a strong Allee effect is present. A locally stable cycle surrounds a locally stable equilibrium point in the first quadrant as well as an unstable limit cycle, which serves as their separatrix in the phase plane; i.e., there is a range for population sizes for which there exists both autoregulation for the predator-prey system and prey and predator populations approach equilibrium, depending upon the population size. The third ω-limit set is the origin, stating that extinction is always possible in the presence of a strong Allee effect in our model. The fourth ω-limit set is a stable focus surrounded by the third limit cycle mentioned above.

In order to illustrate the result stated in Theorem 8 and the *multistability*, Figure 10 shows a numerical simulation for system (5), which was affected with the software MATLAB [30]. Here $a = 3$, $b = 1$, $k = 6.79211$, $M = 4.07697$, $c = 4.05116$, and $L = L_+ + 0.000001 \approx -0.498898$ (see (12)). Initial conditions at $(x(0), y(0))$ are $(0.8, 0.8)$, $(1.01, 1.01)$, $(2, 2)$, $(2.2, 2.2)$, $(1, 0.95)$, and $(0.4, 0.1)$.

In Figure 10 it is clear that a limit cycle lies between the orbits by the points $(2, 2)$ and $(2.2, 2.2)$, respectively, since the ω-limit of $(2, 2)$ is the origin, but the ω-limit of $(2.2, 2.2)$ is the stable focus located in the diagonal near $x = 2.5$. On the other hand, the two infinitesimal limit cycles are surrounding the singular point $(1, 1)$, but they are too small for both the scale of this figure and these values of parameters. Hence, we present another numerical simulation in Figure 11 where both infinitesimal limit cycles are more visible. Here $a = 3$, $b = 1$, $k = 6.56102$, $M = 4.16516$, $c = 4.16276$, and $L = L_+ + 0.01 \approx -0.480256$ (see (12)). Initial conditions at $(x(0), y(0))$ are $(0.6, 0.6)$, $(0.8, 0.8)$, and $(0.95, 0.95)$.

Behaviors of the three positive oriented orbits in Figure 11 imply that, between two of them, attracting and repelling limit cycles exist. Furthermore, it can be easily verified that, for these values of parameters, the divergence of vector field (5) in singu-

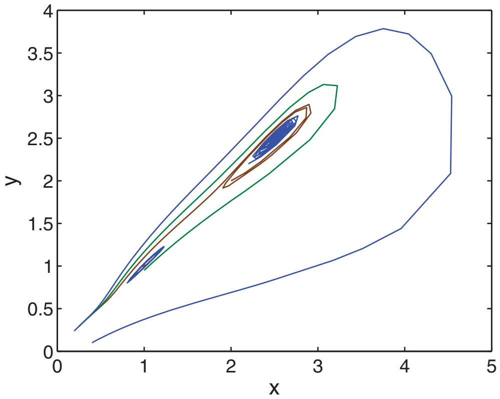


FIG. 10. *Numerical simulation of the results of Theorem* 8.

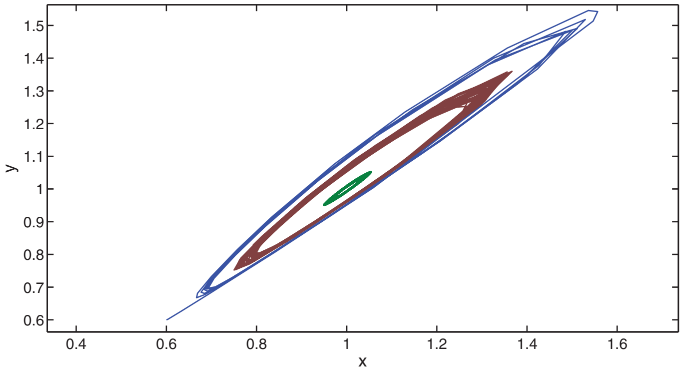


FIG. 11. *Numerical simulation of the infinitesimal limit cycles.*

larity $(1, 1)$ is negative, and, consequently, this equilibrium point is a local attractor and the innermost limit cycle is the unstable one.

Several natural predator-prey communities have been studied, each one of them featuring an ecologically stable cycle, that is, a periodic orbit that must be somewhat insensitive to the perturbations outer to the interaction.

In this work, we have answered partially one of the almost impossible projects proposed in [11] which is the following: Find a predator-prey or other interacting system in nature, or construct one in the laboratory, with at least two ecologically stable cycles.

This shows the ecological relevance of the existence of multiple limit cycles in predation models and the importance of our result, which should serve for the outlined problem to be actually feasible in a biological lab with appropriate little creatures [11].

REFERENCES

[1] P. Aguirre, E. González-Olivares, and E. Sáez, *Two limit cycles in a Leslie–Gower predator-prey model with additive Allee effect*, Nonlinear Anal. Real World Appl., to appear.

[2] A. Andronov, E. Leontovich, I. Gordon, and A. Maier, *Theory of Bifurcations of Dynamical Systems on a Plane,* Israel Program for Scientific Translations, Jerusalem, 1971.

[3] D. Arrowsmith and C. Place, *An Introduction to Dynamical Systems*, Cambridge University Press, New York, 1990.

[4] D. K. Arrowsmith and C. M. Place, *Dynamical Systems. Differential Equations, Maps and Chaotic Behaviour,* Chapman and Hall, London, 1992.

[5] J. Bascompte, *Extinction thresholds: Insights from simple models*, Ann. Zool. Fennici, 40 (2003), pp. 99–114.

[6] L. Berec, E. Angulo, and F. Courchamp, *Multiple Allee effects and population management,* Trends Ecology Evol., 22 (2007), pp. 185–191.

[7] T. R. Blows and N. G. Lloyd, *The number of limit cycles of certain polynomial differential equations,* Proc. Roy. Soc. Edinburgh Sect. A, 98 (1984), pp. 215–239.

[8] D. S. Boukal and L. Berec, *Single-species models and the Allee effect: Extinction boundaries, sex ratios and mate encounters,* J. Theoret. Biol., 218 (2002), pp. 375–394.

[9] C. W. Clark, *Mathematical Bioeconomic. The Optimal Management of Renewable Resources*, 2nd ed., John Wiley and Sons, New York, 1990.

[10] C. W. Clark, *The Worldwide Crisis in Fisheries: Economic Models and Human Behavior,* Cambridge University Press, New York, 2006.

[11] C. S. Coleman, *Hilbert's* 16*th problem: How many cycles?* in Differential Equation Models, Modules Appl. Math. 1, M. Braun, C. S. Coleman, D. A. Drew, and W. F. Lucas, eds., Springe, New York, 1983.

[12] J. B. Collings, *Bifurcation and stability analysis of a temperature-dependent mite predator-prey interaction model incorporating a prey refuge*, Bull. Math. Biol., 57 (1995), pp. 63–76.

[13] F. Courchamp, T. Clutton-Brock, and B. Grenfell, *Inverse dependence and the Allee effect*, Trends Ecology Evol., 14 (1999), pp. 405–410.

[14] F. Courchamp, B. T. Grenfell, and T. H. Clutton-Brock, *Impact of natural enemies on obligately cooperative breeders,* Oikos, 91 (2000), pp. 311–322.

[15] A. M. De Roos and L. Persson, *Size-dependent life-history traits promote catastrophic collapses of top predators,* Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 12907–12912.

[16] B. Dennis, *Allee effects: Population growth, critical density and the chance of extinction*, Natur. Resource Modeling, 3 (1989), pp. 481–538.

[17] F. Dumortier, *Singularities of Vector Fields*, Monogr. Mat. 32, IMPA, Rio de Janeiro, 1978.

[18] R. Etienne, B. Wertheim, L. Hemerik, P. Schneider, and J. Powell, *The interaction between dispersal, the Allee effect and scramble competition affects population dynamics*, Ecol. Model., 148 (2002), pp. 153–168.

[19] J. D. Flores, J. Mena-Lorca, B. González-Yañez, and E. González-Olivares, *Consequences of depensation in a Smith's bioeconomic model for open-access fishery,* in Proceedings of the 2006 International Symposium on Mathematical and Computational Biology BIOMAT 2006, Manaus, Brazil, 2007, pp. 219–232.

[20] H. I. Freedman and G. S. K. Wolkowicz, *Predator-prey systems with group defence: The paradox of enrichment revisted,* Bull. Math. Biol., 8 (1986), pp. 493–508.

[21] J. Gascoigne and R. N. Lipcius, *Allee effects in marine systems*, Marine Ecology Prog. Ser., 269 (2004), pp. 49–59.

[22] J. Gascoigne and R. N. Lipcius, *Allee effects driven by predation,* J. Appl. Ecology, 41 (2004), pp. 801–810.

[23] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[24] I. Hanski, L. Hansson, and H. Henttonen, *Specialist predators, generalist predators and the microtine rodent cycle,* J. Animal Ecology, 60 (1991), pp. 353–367.

[25] I. Hanski, H. Hentonnen, E. Korpimaki, L. Oksanen, and P. Turchin, *Small-rodent dynamics and predation,* Ecology, 82 (2001), pp. 1505–1520.

[26] X. Huang, Y. Wang, and L. Zhu, *One and three limit cycles in a cubic predator-prey system,* Math. Methods Appl. Sci., 30 (2007), pp. 501–511.

[27] A. Korobeinikov, *A Lyapunov function for Leslie-Gower predator-prey models,* Appl. Math. Lett., 14 (2001), pp. 697–699.

[28] Y. Li and D. Xiao, *Bifurcations of a predator-prey system of Holling and Leslie types,* Chaos Solitons Fractals, 34 (2007), pp. 606–620.

[29] D. Luo, X. Wang, D. Zhu, and M. Han, *Bifurcation Theory and Methods of Dynamical Systems,* Adv. Ser. Dyn. Syst. 15, World Scientific, Singapore, 1997.

[30] *MATLAB: The Language of Technical Computing, Using MATLAB (Version* 7.0), MathWorks, Natwick, MA, 2004.

[31] R. M. May, *Stability and Complexity in Model Ecosystems*, 2nd ed., Princeton University Press, Princeton, NJ, 2001.

[32] J. Mena-Lorca, E. González-Olivares, and B. González-Yañez, *The Leslie-Gower predator-prey model with Allee effect on prey: A simple model with a rich and interesting dynamics,* in Proceedings of the 2006 International Symposium on Mathematical and Computational Biology BIOMAT 2006, Manaus, Brazil, 2007, pp. 105–132.

[33] E. Sáez and E. González-Olivares, *Dynamics of a predator-prey model*, SIAM J. Appl. Math., 59 (1999), pp. 1867–1878.

[34] P. A. Stephens and W. J. Sutherland, *Consequences of the Allee effect for behaviour, ecology and conservation*, Trends Ecology Evol., 14 (1999), pp. 401-405.

[35] F. Takens, *Unfoldings of certain singularities of vector fields: Generalized Hopf bifurcations*, J. Differential Equations, 14 (1973), pp. 476–493.

[36] R. J. Taylor, *Predation*, Chapman and Hall, New York, 1984.

[37] P. Turchin, *Complex Population Dynamics. A Theoretical/Empirical Synthesis,* Monogr. Population Biol., 35, Princeton University Press, Princeton, NJ, 2003.

[38] X. Wang, G. Liang, and F.-Z. Wang, *The competitive dynamics of populations subject to an Allee effect*, Ecol. Model., 124 (1999), pp. 130–168.

[39] G. S. W. Wolkowicz, *Bifurcation analysis of a predator-prey system involving group defense,* SIAM J. Appl. Math., 48 (1988), pp. 592–606.

[40] G. S. K. Wolkowicz, H. Zhu, and S. A. Campbell, *Bifurcation analysis of a predator-prey system with nonmonotonic functional response,* SIAM J. Appl. Math., 63 (2003), pp. 636–682.

[41] S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, Addison–Wesley, Longman, Boston, 1988

[42] D. Xiao and S. Ruan, *Global analysis in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 61 (2001), pp. 1445–1472.

[43] D. Xiao and S. Ruan, *Bifurcations in a predator-prey system with group defense,* Int. J. Bifurcation and Chaos, 11 (2001), pp. 2123–2131.

[44] D. Xiao and H. Zhu, *Multiple focus and Hopf bifurcations in a predator-prey system with nonmonotonic functional response,* SIAM J. Appl. Math., 66 (2006), pp. 802–819.

[45] D. Xiao and K. F. Zhang, *Multiple bifurcations of a predator-prey system,* Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 417–433.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1263–1276

DOUBLE-DIFFUSIVE CONVECTION IN A POROUS LAYER IN THE PRESENCE OF VIBRATION*

NATALIA STRONG†

Abstract. The present paper examines the effect of vertical harmonic vibration on the onset of convection in an infinite horizontal layer of a binary fluid mixture saturating a porous medium. A constant temperature and concentration distribution are assigned on the rigid boundaries, so that there exist vertical temperature and concentration gradients. The mathematical model is described by equations of filtration convection in the Darcy–Oberbeck–Boussinesq approximation. The linear stability analysis for the quasi-equilibrium solution is performed using Floquet theory. Employing the method of continued fractions allows derivation of the dispersion equation for the Floquet exponent in an explicit form. The neutral curves of the Rayleigh number versus horizontal wave number are constructed for the three types of instability modes: synchronous, subharmonic, and complex conjugate. Asymptotic formulas for these curves are derived for large values of vibration frequency using the method of averaging. It is shown that, at some finite frequencies of vibration, there exist regions of parametric instability. Investigations carried out in the paper demonstrate that, depending on the governing parameters of the problem, vertical vibration can significantly affect the stability of the system by increasing or decreasing its susceptibility to convection. In addition, even in the presence of vibration, the onset of convection in the system is affected by variations in the concentration of the solute in the mixture.

Key words. double-diffusive convection, instability, porous media, resonance, vibration

AMS subject classifications. 76E15, 76S05, 35Q35

DOI. 10.1137/060674776

1. Introduction.

Natural double-diffusive convection in fluid-saturated porous media has received much attention during the last few decades and has practical importance in many fields such as geophysics, oceanography, ecology, chemistry, and metallurgy. Specific areas of application range from the flow of groundwater to oil recovery, underground storage of waste products, food processing, and building insulation.

Double-diffusive convection in an infinite horizontal layer of a porous medium was first investigated by Nield using linear stability analysis [1] and was later extended by Taunton and Lightfoot [2]. Rudraiah, Srimani, and Friedrich [3] applied nonlinear stability analysis to the case of a porous layer with isothermal and isosolutal boundaries. Further works on double-diffusive convection in porous media include Brand and Steinberg [4], Murray and Chen [5], and Mamou and Vasseur [6]. An extensive review of the literature on natural convection in fluid-saturated porous media may be found in the fundamental monograph by Nield and Bejan [7].

In addition to free convection in porous media, an important class of problems involves convective instability in the presence of time-dependent body forces, one of which is vibration. The time-dependent gravitational field is of great interest, for example, in space laboratory experiments, crystal growth, petroleum production, and large-scale atmospheric convection.

*Received by the editors November 12, 2006; accepted for publication (in revised form) October 27, 2008; published electronically February 4, 2009. An extended abstract of this paper was published in the proceedings of the workshop *Women in Mathematics: The Legacy of Ladyzhenskaya and Oleinik*, MSRI, Berkeley, CA, 2006.

http://www.siam.org/journals/siap/69-5/67477.html

†Department of Mathematics, University of Missouri, Columbia, MO 65211 (natalya@math.missouri.edu).

Although much has been published on convection in the presence of vibration in fluids, only limited attention has been given to this phenomenon in porous media. The onset of convection in a region of fluid saturating a porous medium subjected to high-frequency vibration of arbitrary direction has been examined by Zen'kovskaya [8] and Zen'kovskaya and Rogovenko [9] using the averaging method. It was shown that the direction of vibration has a significant effect on the stability of the system. Malashetty and Padmavathi [10] performed asymptotical analysis of the linear stability of a horizontal fluid-saturated porous layer heated from below for the case of small-amplitude gravity modulation. Bardan and Mojtabi [11] studied, numerically and analytically, convection in a rectangular saturated porous cavity heated from below and subjected to high-frequency vibration. They concluded that high-frequency vibration moves the onset of convection toward higher values of the Rayleigh number. Jounet and Bardan [12] generalized the work described above to the case of a binary mixture saturating a porous medium. They found that high-frequency vibration can delay or speed up the onset of convection depending on the governing parameters of the problem. Govender [13] analytically investigated convection in a porous layer heated from below for the case of low-amplitude vibration and showed that increasing the frequency of vibration stabilizes the convection.

In the present paper, we generalize the work on the pure fluid case [14] to a binary fluid mixture saturating a porous medium. Specifically, we investigate the effect of vertical vibration of arbitrary frequency and amplitude on the onset of convection in a horizontal layer of a porous medium saturated by a binary fluid mixture. Previous studies of this problem were restricted to the case of high frequency and/or small amplitude of vibration due to the limitations of the methods used. This paper's novel application of the method of continued fractions to the above-described problem eliminates these restrictions and enables consideration of vibration of arbitrary frequency and amplitude. Moreover, employing the method of continued fractions allows derivation of the dispersion equation for the Floquet exponent in an explicit form. This dispersion equation is used to find the critical values of parameters and to construct the neutral curves corresponding to the three types of transition to instability: the synchronous, subharmonic, and complex conjugate modes. In the case of high-frequency vibration, the system is investigated using the method of averaging, which reduces the dispersion equation for the Floquet exponent to a cubic equation. Comparing the results obtained by the two methods described above allows us to find the range of vibration parameter values for which the method of continued fractions can be replaced by the method of averaging in numerical computations.

The structure of the paper is as follows. Section 2 describes the mathematical model and the mechanical quasi-equilibrium solution to the system. Linear stability analysis of this quasi-equilibrium solution is performed in section 3 following Floquet theory and employing the method of continued fractions. The case of high-frequency vibration of the layer is considered in section 4 using the averaging method. Section 5 presents and discusses the results of the numerical computations.

2. Problem description and basic equations. We consider an infinite horizontal layer of a porous medium saturated by a binary mixture of nonreacting components. The layer and its boundaries are subjected to vertical harmonic vibration. We assume that the fluid component of the mixture is viscous and incompressible, the porous medium is homogeneous and isotropic, and the boundaries are rigid and impermeable, with slip allowed. Constant and uniform temperature and concentration distributions are specified at the boundaries, so that there exist vertical temperature and concentration gradients.

The equation of state has the following form:

$$\rho = \rho_0[1 - \beta(T - T_0) + \beta_c(C - C_0)],$$

where ρ is the density of the fluid mixture, T is the temperature of the fluid mixture and the porous medium, and C is the concentration of the heavier component of the fluid mixture; $\rho_0 = \rho(T_0, C_0)$ is the density at some reference temperature T_0 and concentration C_0, β is the coefficient of the thermal expansion of the fluid mixture, and β_c is the coefficient of the concentration expansion of the fluid mixture ($\beta > 0$ and $\beta_c > 0$). We assume that $T_0 = 0$, $C_0 = 0$, and variations of the temperature and concentration in the fluid are sufficiently small.

The basic governing equations for the problem are the momentum equation for a porous medium, the energy conservation equation, the mass conservation equation, and the continuity equation. We assume that the Oberbeck–Boussinesq approximation is valid, which implies that the density variation is included only in the gravitational term of the momentum equation. In the Cartesian coordinate system inflexibly fixed to the oscillating horizontal layer, with the z-axis directed vertically upward and the origin at the lower plate, the governing equations have the following form:

$$\frac{1}{\varphi}\frac{\partial \mathbf{v}}{\partial t} = -\frac{1}{\rho}\nabla p - \frac{\nu}{K}\mathbf{v} + g(t)\left(\beta T - \beta_c C\right)\mathbf{k}, \tag{2.1}$$

$$\varkappa\frac{\partial T}{\partial t} + \mathbf{v}\cdot\nabla T = \chi\nabla^2 T, \tag{2.2}$$

$$\varphi\frac{\partial C}{\partial t} + \mathbf{v}\cdot\nabla C = D_m\nabla^2 C, \tag{2.3}$$

$$\nabla\cdot\mathbf{v} = 0, \tag{2.4}$$

where $\mathbf{v} = (v_1, v_2, v_3)$ is the relative filtration velocity of the fluid mixture, p is the convective pressure, φ is the porosity of the medium, ν is the kinematic viscosity of the fluid mixture, K is the intrinsic permeability of the porous medium, $\mathbf{k}$ is the unit vector directed upward, $\varkappa$ is the heat capacity ratio (porous medium versus fluid), χ is the thermal diffusivity of the porous medium, and D_m is the mass diffusivity of the porous medium.

Due to the vertical vibration of the layer, the gravitational field $g(t)$ in the momentum equation consists of two parts: $g(t) = g_0 + g_e(t)$. The first term g_0 is the steady acceleration due to the static gravity. The second term $g_e(t) = \frac{A}{\varphi}\,\Omega^2 f''(\tau)$ represents the vibrational acceleration and implies that the vertical motion of the layer is described by the formula $z = Af(\tau)$. Here $\tau = \Omega t$, A is the amplitude, Ω is the frequency of vibration, and $f(\tau)$ is a 2π-periodic function with zero 2π-average:

$$\langle f\rangle := \frac{1}{2\pi}\int_0^{2\pi} f(\tau)d\tau = 0.$$

We consider the boundary conditions

$$z = 0: \qquad v_3 = 0, \qquad T = T_1, \qquad C = C_1, \tag{2.5}$$

$$z = h: \qquad v_3 = 0, \qquad T = T_2, \qquad C = C_2, \tag{2.6}$$

which imply constant and uniform distribution of temperature and concentration along the rigid and impermeable walls, where slip is allowed. No assumptions are made about the relative size of the temperature boundary conditions (T_1 versus T_2) or concentration boundary conditions (C_1 versus C_2), meaning that the temperature and concentration gradients can be of any sign.

The system (2.1)–(2.6) has a solution corresponding to the quasi-equilibrium basic state:

$$\mathbf{v}^0 = 0, \quad T^0 = T_1 - \frac{1}{h}(T_1 - T_2)z, \quad C^0 = C_1 - \frac{1}{h}(C_1 - C_2)z,$$
$$p^0 = \rho g(t)\left[\left(\beta T_1 - \beta_c C_1\right)z - \frac{1}{2h}\Big(\beta(T_1 - T_2) - \beta_c(C_1 - C_2)\Big)z^2\right]. \tag{2.7}$$

To analyze the stability of this basic state by using the linearization method, we introduce small perturbations of the variables:

$$\mathbf{v} = \mathbf{v}^0 + \mathbf{u}, \qquad p = p^0 + q, \qquad T = T^0 + \theta, \qquad C = C^0 + S. \tag{2.8}$$

The following scales are used to nondimensionalize the variables:

$$(x, t, \mathbf{v}, p, T, C) \to \left(h, \frac{h^2}{\nu}, \frac{\nu}{h}, \frac{\rho\nu^2}{K}, ah, bh\right),$$

where $a = (T_1 - T_2)/h$ and $b = (C_1 - C_2)/h$ are the quasi-equilibrium temperature and concentration gradients, respectively. The dimensionless linearized system has the following form (with notation for the dimensionless variables being the same as for corresponding dimensional variables):

$$c\frac{\partial \mathbf{u}}{\partial t} = -\nabla q - \mathbf{u} + \left(1 + \eta f''(\tau)\right)\left(\mathrm{Gr}\,\theta - \mathrm{Gr}_c S\right)\mathbf{k}, \tag{2.9}$$
$$\varkappa\frac{\partial \theta}{\partial t} - u_3 = \frac{1}{\mathrm{Pr}}\nabla^2\theta, \tag{2.10}$$
$$\varphi\frac{\partial S}{\partial t} - u_3 = \frac{1}{\mathrm{Sc}}\nabla^2 S, \tag{2.11}$$
$$\nabla \cdot \mathbf{u} = 0, \tag{2.12}$$

where $c = \frac{K}{\varphi h^2}$, $\mathrm{Gr} = \frac{\beta a h^2 g_0 K}{\nu^2}$ is the thermal Grashoff number, $\mathrm{Gr_c} = \frac{\beta_c b h^2 g_0 K}{\nu^2}$ is the concentration Grashoff number, $\mathrm{Pr} = \frac{\nu}{\chi}$ is the Prandtl number, $\mathrm{Sc} = \frac{\nu}{D_m}$ is the Schmidt number, and $\eta = \frac{A\Omega^2}{\varphi g_0}$ and $\omega = \frac{\Omega h^2}{\nu}$ are the nondimensional amplitude and frequency of vibration, respectively. Dimensionless boundary conditions are

$$z = 0: \qquad u_3 = \theta = S = 0, \tag{2.13}$$
$$z = 1: \qquad u_3 = \theta = S = 0. \tag{2.14}$$

3. Derivation of the dispersion equation for the Floquet exponent $\boldsymbol{\sigma}$. In order to eliminate pressure and the horizontal components of velocity from the system (2.9)–(2.14), we apply the curl operator twice to (2.9). The z-component of the resulting equation has the following form:

$$\left(c\frac{\partial}{\partial t} + 1\right)\nabla^2 u_3 = \Big(1 + \eta f''(\tau)\Big)\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Big[\mathrm{Gr}\,\theta - \mathrm{Gr}_c S\Big]. \tag{3.1}$$

Now we consider the system of equations (2.10), (2.11), and (3.1) with the boundary conditions (2.13)–(2.14). The x- and y-variables can be separated from this system by introducing the following representation for the perturbations of vertical velocity, temperature, and concentration:

$$\begin{bmatrix} u_3 \\ \theta \\ S \end{bmatrix} (x, y, z, t) = \begin{bmatrix} \widetilde{u_3} \\ \widetilde{\theta} \\ \widetilde{S} \end{bmatrix} (z, t)\ e^{i(\alpha_1 x + \alpha_2 y)}.$$

Substituting this representation into the system of equations (2.10), (2.11), and (3.1) reduces it to the following form:

$$\left(c\frac{\partial}{\partial t} + 1\right) \left[\frac{\partial^2}{\partial z^2} - \alpha^2\right] \widetilde{u_3} = -\alpha^2 \Big(1 + \eta f\,''(\tau)\Big) \left[\mathrm{Gr}\ \widetilde{\theta} - \mathrm{Gr}_c\ \widetilde{S}\right], \tag{3.2}$$

$$\varkappa \frac{\partial \widetilde{\theta}}{\partial t} - \widetilde{u_3} = \frac{1}{\mathrm{Pr}} \left(\frac{\partial^2}{\partial z^2} - \alpha^2\right) \widetilde{\theta}, \tag{3.3}$$

$$\varphi \frac{\partial \widetilde{S}}{\partial t} - \widetilde{u_3} = \frac{1}{\mathrm{Sc}} \left(\frac{\partial^2}{\partial z^2} - \alpha^2\right) \widetilde{S}, \tag{3.4}$$

where $\alpha^2 = \alpha_1^2 + \alpha_2^2$ is the square of the overall horizontal wave number. Boundary conditions for this system are

$$z = 0,\ z = 1: \qquad \widetilde{u_3} = \widetilde{\theta} = \widetilde{S} = 0. \tag{3.5}$$

Now we separate the z-variable from the system (3.2)–(3.4) using the following representation:

$$\begin{bmatrix} \widetilde{u}_3 \\ \widetilde{\theta} \\ \widetilde{S} \end{bmatrix} (z, t) = \begin{bmatrix} \widehat{u}_3 \\ \widehat{\theta} \\ \widehat{S} \end{bmatrix} (t) \sin(\pi \ell z), \qquad \ell = 1, 2, 3, \dots.$$

After the substitutions $t = \widehat{t}\sqrt{\mathrm{Pr}\,\varkappa c}$ and $\omega = \widehat{\omega}/\sqrt{\mathrm{Pr}\,\varkappa c}$, we obtain a system of ODEs with periodic coefficients:

$$\frac{c}{r}\frac{d\widehat{u}_3}{d\widehat{t}} = -\widehat{u}_3 + \frac{\alpha^2}{m^2}\Big(1 + \eta f\,''(\tau)\Big)\left(\mathrm{Gr}\ \widehat{\theta} - \mathrm{Gr}_c \widehat{S}\right), \tag{3.6}$$

$$\frac{\varkappa}{r}\frac{d\widehat{\theta}}{d\widehat{t}} = \widehat{u}_3 - \frac{m^2}{\mathrm{Pr}}\widehat{\theta}, \tag{3.7}$$

$$\frac{\varphi}{r}\frac{d\widehat{S}}{d\widehat{t}} = \widehat{u}_3 - \frac{m^2}{\mathrm{Sc}}\widehat{S}, \tag{3.8}$$

where $m^2 = \alpha^2 + (\pi\ell)^2$ and $r = \sqrt{\mathrm{Pr}\,\varkappa c}$. For the rest of the paper, we omit the hat in $\widehat{t}$ and $\widehat{\omega}$ and assume that $f(\tau) = \cos\omega t$ in (3.6).

Following Floquet theory (see [15]), we search for the solution to the system (3.6)–(3.8) in the form of a Fourier series:

$$\begin{bmatrix} \widehat{u}_3 \\ \widehat{\theta} \\ \widehat{S} \end{bmatrix} (t) = e^{\sigma t} \sum_{n=-\infty}^{+\infty} \begin{bmatrix} w_n \\ \theta_n \\ S_n \end{bmatrix} e^{in\omega t}. \tag{3.9}$$

The Floquet exponent σ needs to be chosen in such a way that solution (3.9) is nonzero. The set of all possible values of σ defines the *Floquet spectrum* of the system

(3.6)–(3.8). The behavior of solution (3.9), and therefore the stability of the quasi-equilibrium basic state (2.7) of the original system, is determined by the distribution of the Floquet spectrum with respect to the imaginary axis in the complex σ-plane. If the whole spectrum is located in the left half-plane, the basic state (2.7) is asymptotically stable. If at least one point of the spectrum is located in the right half-plane, the basic state (2.7) is unstable. The points of the spectrum where $\operatorname{Re}(\sigma) = 0$ correspond to the *neutral curves* (or *marginal curves*) in the parameter space, which separate regions of stability and instability.

Substituting representation (3.9) into the system (3.6)–(3.8) and collecting coefficients of $e^{in\omega t}$ for each n, we obtain an infinite system of linear algebraic equations for the Fourier coefficients w_n, θ_n, S_n:

$$\begin{aligned}\frac{2m^2}{\alpha^2\eta}\left[c(\sigma+in\omega)+1\right]w_n &= \mathrm{Gr}_c\left(S_{n-1}+S_{n+1}-\frac{2}{\eta}S_n\right)\\ &\quad-\mathrm{Gr}\left(\theta_{n-1}+\theta_{n+1}-\frac{2}{\eta}\theta_n\right),\\ w_n &= \left[\varkappa(\sigma+in\omega)+\frac{m^2}{\mathrm{Pr}}\right]\theta_n,\\ w_n &= \left[\varphi(\sigma+in\omega)+\frac{m^2}{\mathrm{Sc}}\right]S_n, \qquad n=0,\pm1,\pm2,\dots.\end{aligned}$$

By eliminating the variables w_n and S_n from this system, we transform it into an infinite tridiagonal system of linear algebraic equations for the coefficients θ_n:

$$M_n\theta_n + q_{n-1}\theta_{n-1} + q_{n+1}\theta_{n+1} = 0, \qquad n = 0, \pm1, \pm2, \dots, \tag{3.10}$$

$$\text{with} \qquad M_n = (\sigma+\ in\omega)^2 + \left(\frac{m^2}{P}+P\right)(\sigma+\ in\omega) + \left(m^2 - \frac{2q_n}{\eta}\right), \tag{3.11}$$

$$q_n = \frac{\alpha^2\eta}{2m^2}\left[\mathrm{Ra} - \mathrm{Rs}\frac{(\sigma+\ in\omega)+\dfrac{m^2}{P}}{\mathrm{L}(\sigma+\ in\omega)+\dfrac{m^2}{P}}\right], \quad \mathrm{L} = \mathrm{Le}\,\frac{\varphi}{\varkappa}, \quad P = \sqrt{\frac{\mathrm{Pr}\,\varkappa}{c}}\,. \tag{3.12}$$

Here $\mathrm{Ra} = \mathrm{Gr}\cdot\mathrm{Pr}$ is the thermal Rayleigh number, $\mathrm{Rs} = \mathrm{Gr_c}\cdot\mathrm{Sc}$ is the concentration Rayleigh number, and $\mathrm{Le} = \mathrm{Sc}/\mathrm{Pr}$ is the Lewis number.

We use the method of continued fractions to solve the system (3.10) (see Meshalkin and Sinai [16], followed by Yudovich [17], Markman and Yudovich [18], and Yudovich et al. [19]). Substituting $\zeta_n = \frac{\theta_{n-1}}{\theta_n}$ (in the assumption that $\theta_n \neq 0$, which is verified in [18]) transforms the system (3.10) into

$$M_n + q_{n-1}\zeta_n + \frac{q_{n+1}}{\zeta_{n+1}} = 0, \qquad n = 0, \pm1, \pm2, \dots,$$

which can be used to derive two recurrence relations for the unknown ζ_n:

$$\begin{aligned}\zeta_n &= -\frac{1}{q_{n-1}}\left(M_n + \frac{q_{n+1}}{\zeta_{n+1}}\right),\\ \zeta_n &= \frac{-q_n}{M_{n-1}+\zeta_{n-1}\,q_{n-2}}.\end{aligned}$$

Each of these two relations for ζ_n yields a corresponding continued fraction expression:

$$\zeta_n = -\frac{M_n}{q_{n-1}} - \cfrac{q_{n+1}/q_{n-1}}{-\cfrac{M_{n+1}}{q_n} - \cfrac{q_{n+2}/q_n}{-\cfrac{M_{n+2}}{q_{n+1}} - \cfrac{q_{n+3}/q_{n+1}}{-\cfrac{M_{n+3}}{q_{n+2}} - \cdots}}}\,, \tag{3.13}$$

$$\zeta_n = \cfrac{q_n}{-M_{n-1} - \cfrac{q_{n-1}q_{n-2}}{-M_{n-2} - \cfrac{q_{n-2}q_{n-3}}{-M_{n-3} - \cdots}}}\,. \tag{3.14}$$

The right-hand sides of relations (3.13) and (3.14) are equal to each other for any integer value of n. Therefore, assigning $n = 0$, we obtain the dispersion equation for the Floquet exponent σ in the explicit form

$$M_0 - \cfrac{q_0\ q_1}{M_1 - \cfrac{q_1\ q_2}{M_2 - \cfrac{q_2\ q_3}{M_3 - \cdots}}} = \cfrac{q_0\ q_{-1}}{M_{-1} - \cfrac{q_{-1}q_{-2}}{M_{-2} - \cfrac{q_{-2}q_{-3}}{M_{-3} - \cdots}}}. \tag{3.15}$$

Equation (3.15) can be used to determine the Floquet spectrum of the system (3.6)–(3.8) with all the values of parameters being fixed. Alternatively, (3.15) can be used to determine the critical values of parameters, corresponding to transition from stability to instability of the basic state (2.7), and to construct the corresponding neutral curves.

Convergence of continued fractions from (3.15) was verified numerically. Moreover, the sufficient condition for the absolute convergence of the continued fraction in the left-hand side of (3.15) is

$$|M_n| \geq 1 + |q_{n-1}\, q_n|, \qquad n = 1, 2, 3, \ldots. \tag{3.16}$$

It follows from (3.11) and (3.12) that $|M_n|$ is proportional to n^2 and $|q_n|$ is proportional to $O(1)$. Hence, there exists some N such that condition (3.16) is satisfied for any $n \geq N$. This fact guarantees the conditional convergence of the continued fraction in the left-hand side of (3.15) (see [20]). The same argument (with n being replaced by $-n$) can be used to prove conditional convergence of the continued fraction in the right-hand side of (3.15).

Now we consider two particular cases for which (3.15) simplifies to a real form:

1. For the case $\sigma = 0$, corresponding to synchronous modes with period $2\pi/\omega$, the expressions for M_n and q_n are simplified so that

$$M_{-n} = \overline{M_n} \qquad \text{and} \qquad q_{-n} = \overline{q_n}\,.$$

Therefore, dispersion equation (3.15) for the case $\sigma = 0$ transforms into

$$\operatorname{Re}\left(\cfrac{q_0\ q_1}{M_1 - \cfrac{q_1\ q_2}{M_2 - \cfrac{q_2\ q_3}{M_3 - \cdots}}} \right) = \frac{M_0}{2}. \tag{3.17}$$

2. For the case $\sigma = \frac{i\omega}{2}$, corresponding to subharmonic modes with period $4\pi/\omega$, the expressions for M_n and q_n are simplified so that
$$M_{-n} = \overline{M_{n-1}} \qquad \text{and} \qquad q_{-n} = \overline{q_{n-1}}\,.$$
Therefore, dispersion equation (3.15) for the case $\sigma = \frac{i\omega}{2}$ transforms into
$$\left| M_0 - \cfrac{q_0\ q_1}{M_1 - \cfrac{q_1\ q_2}{M_2 - \cfrac{q_2\ q_3}{M_3 - \cdots}}} \right|^2 = q_0^2. \tag{3.18}$$

Transcendental equations (3.17) and (3.18) were investigated numerically to obtain the neutral curves of the Rayleigh number Ra versus the horizontal wavenumber α for the synchronous and subharmonic modes. In addition to the synchronous and subharmonic modes, which also occur in the case of a pure fluid saturating a porous medium, instability in the binary mixture case can occur via a complex conjugate mode. In order to construct the neutral curves $\mathrm{Ra}(\alpha)$ for complex conjugate modes, we solve a complex equation (3.15) for the two variables: the thermal Rayleigh number and the frequency of neutral oscillations. Results of the numerical computations are presented in section 5.

4. The case of a rapidly oscillating external force. In the limiting case of high-frequency vibration, we apply the Krylov–Bogoliubov averaging method to investigate the stability of the basic state (2.7) of the original system (see [21], [8], [9]). First, we eliminate the variable $\widehat{u}_3$ from the system (3.6)–(3.8) (with the hat omitted in $\widehat{t}$ and $\widehat{\omega}$) and reduce it to the form
$$\begin{aligned}\widehat{\theta}\,''(t) + \left(\frac{m^2}{P} + P\right)\widehat{\theta}\,'(t) + \left(m^2 - \frac{\alpha^2}{m^2}\mathrm{Ra}\,(1 - \eta\cos\omega t)\right)\widehat{\theta}(t)\\ + \frac{\alpha^2}{m^2}\,\mathrm{Rs}\ \mathrm{Le}^{-1}\,(1 - \eta\cos\omega t)\,\widehat{S}(t) = 0,\end{aligned} \tag{4.1}$$
$$\widehat{\theta}\,'(t) + \frac{m^2}{P}\widehat{\theta}(t) = \frac{\varphi}{\varkappa}\ \widehat{S}\,'(t) + \frac{m^2}{P}\mathrm{Le}^{-1}\widehat{S}(t). \tag{4.2}$$

Now, assuming $\omega \to \infty$, we represent the nondimensional amplitude of vibration in the form $\eta = B\omega$, where $B = \frac{b\,\nu}{\varphi\,g_0 h^2} = O(1)$ is the nondimensional amplitude of vibration velocity. Following the averaging method, we introduce a new "fast" time variable $\tau = \omega t$ and decompose the solution to the system (4.1)–(4.2) in the following way:
$$\begin{bmatrix} \widehat{\theta} \\ \widehat{S} \end{bmatrix}(t) \sim \begin{bmatrix} \bar{\theta} \\ \bar{S} \end{bmatrix}(t) + \frac{1}{\omega}\begin{bmatrix} \widetilde{\theta} \\ \widetilde{S} \end{bmatrix}(\tau, t), \tag{4.3}$$
where $\big(\bar{\theta}(t),\ \bar{S}(t)\big)$ is a slowly varying part, which is a function of the "slow" time t only, and $\big(\widetilde{\theta}(\tau,t),\ \widetilde{S}(\tau,t)\big)$ is a rapidly oscillating part, which is a function of τ and t. The functions $\widetilde{\theta}(\tau,t)$ and $\widetilde{S}(\tau,t)$ are assumed to be 2π-periodic in τ and have zero mean with respect to τ. Substituting representation (4.3) into the system (4.1)–(4.2)

and retaining the principal ($O(1)$) terms, we obtain a system of equations for the rapidly oscillating part $\big(\widetilde{\theta}(\tau,t),\ \widetilde{S}(\tau,t)\big)$:

$$\widetilde{\theta}_{\tau\tau}(\tau,t) = \frac{\alpha^2}{m^2} B\cos\tau\big(\mathrm{Rs}\,\mathrm{Le}^{-1}\bar{S}(t) - \mathrm{Ra}\,\bar{\theta}(t)\big), \tag{4.4}$$

$$\frac{\varphi}{\varkappa}\,\widetilde{S}_{\tau}(\tau,t) = \widetilde{\theta}_{\tau}(\tau,t). \tag{4.5}$$

Integrating this system with respect to the "fast" time τ (under the assumption that $\bar{\theta}(t)$ and $\bar{S}(t)$ do not depend on τ) gives us expressions for the rapidly oscillating part:

$$\widetilde{\theta}(\tau,t) = \frac{\alpha^2}{m^2} B\cos\tau\left(\mathrm{Ra}\;\bar{\theta}(t) - \mathrm{Rs}\;\mathrm{Le}^{-1}\bar{S}(t)\right), \tag{4.6}$$

$$\widetilde{S}(\tau,t) = \frac{\alpha^2}{m^2}\frac{\varkappa}{\varphi} B\cos\tau\left(\mathrm{Ra}\;\bar{\theta}(t) - \mathrm{Rs}\;\mathrm{Le}^{-1}\bar{S}(t)\right). \tag{4.7}$$

Now we substitute these expressions back into (4.3), then substitute the resulting expressions for $\widehat{\theta}(t)$ and $\widehat{S}(t)$ into the system (4.1)–(4.2), and take the average of this system with respect to the "fast" time τ over the modulation period 2π. This process leads to the system of averaged equations for the slowly varying component $\big(\bar{\theta}(t),\ \bar{S}(t)\big)$:

$$\begin{aligned}\bar{\theta}''(t) + \left(\frac{m^2}{P} + P\right)\bar{\theta}'(t) + \left(m^2 - \frac{\alpha^2}{m^2}\mathrm{Ra}\right)\bar{\theta}(t) + \frac{\alpha^2}{m^2}\;\mathrm{Rs}\;\mathrm{Le}^{-1}\bar{S}(t)\\ + \frac{\alpha^4 B^2}{2m^4}\left(\mathrm{Ra}\,\bar{\theta}(t) - \mathrm{Rs}\;\mathrm{Le}^{-1}\bar{S}(t)\right)\left(\mathrm{Ra} - \frac{\varkappa}{\varphi}\,\mathrm{Rs}\;\mathrm{Le}^{-1}\right) = 0,\end{aligned} \tag{4.8}$$

$$\bar{\theta}'(t) - \frac{\varphi}{\varkappa}\;\bar{S}'(t) = \frac{m^2}{P}\left(\mathrm{Le}^{-1}\;\bar{S}(t) - \bar{\theta}(t)\right). \tag{4.9}$$

This is a system of autonomous equations with constant coefficients. Searching for the solution to this system in the form

$$\begin{bmatrix}\bar{\theta}(t)\\ \bar{S}(t)\end{bmatrix} = \begin{bmatrix}C_1\\ C_2\end{bmatrix} e^{\sigma t}, \quad \text{where} \quad C_1, C_2 = \mathrm{const},$$

yields the characteristic equation for the complex growth rate σ, which determines the stability of the quasi-equilibrium basic state (2.7) in the limiting case of $\omega \to \infty$:

$$b_0\sigma^3 + b_1\sigma^2 + b_2\sigma + b_3 = 0, \tag{4.10}$$

$$\text{where} \quad \mathrm{L} = \mathrm{Le}\,\frac{\varphi}{\varkappa}, \qquad b_0 = \mathrm{L}, \qquad b_1 = \frac{m^2}{P}\,(\mathrm{L}+1) + \mathrm{L}\,P,$$

$$b_2 = \frac{\alpha^4 B^2}{2m^4\mathrm{L}}\,(\mathrm{Rs} - \mathrm{Ra}\,\mathrm{L})^2 + \frac{\alpha^2}{m^2}\,(\mathrm{Rs} - \mathrm{Ra}\;\mathrm{L}) + \left(\frac{m^2}{P} + P\right)\frac{m^2}{P},$$

$$b_3 = \frac{m^2}{P}\left[\frac{\alpha^4 B^2}{2m^4\,\mathrm{L}}\,(\mathrm{Rs} - \mathrm{Ra}\,\mathrm{L})\,(\mathrm{Rs} - \mathrm{Ra}) + \frac{\alpha^2}{m^2}\,(\mathrm{Rs} - \mathrm{Ra}) + m^2\right].$$

For the case of transition to monotonic instability ($\sigma = 0$), the equation for the neutral curves is obtained from the characteristic equation (4.10) and has the form $b_3 = 0$, or

$$\frac{\alpha^4 B^2}{2m^4\,\mathrm{L}}\,(\mathrm{Rs} - \mathrm{Ra}\,\mathrm{L})\,(\mathrm{Rs} - \mathrm{Ra}) + \frac{\alpha^2}{m^2}\,(\mathrm{Rs} - \mathrm{Ra}) + m^2 = 0. \tag{4.11}$$

For the case of transition to oscillatory instability, which occurs via a complex conjugate mode ($\sigma = is,\ s \neq 0$), the equation for the neutral curves is obtained from the characteristic equation (4.10) and has the form $b_0 b_3 - b_1 b_2 = 0$, or

$$\begin{aligned} &\frac{m^2}{P}\mathrm{L}\left[\frac{\alpha^4 B^2}{2m^4\,\mathrm{L}}\,(\mathrm{Rs}-\mathrm{Ra}\,\mathrm{L})\,(\mathrm{Rs}-\mathrm{Ra}) + \frac{\alpha^2}{m^2}\,(\mathrm{Rs}-\mathrm{Ra}) + m^2\right] \\ &\quad = \left(\frac{\alpha^4 B^2}{2m^4\mathrm{L}}\,(\mathrm{Rs}-\mathrm{Ra}\,\mathrm{L})^2 + \frac{\alpha^2}{m^2}\,(\mathrm{Rs}-\mathrm{Ra}\ \mathrm{L}) + \left(\frac{m^2}{P}+P\right)\frac{m^2}{P}\right) \\ &\quad \times \left(\frac{m^2}{P}\,(\mathrm{L}+1) + \mathrm{L}\,P\right). \end{aligned} \tag{4.12}$$

The formula for the square of the frequency s of the neutral oscillations has the form $s^2 = b_2/b_0$, or

$$s^2 = \frac{1}{\mathrm{L}}\left[\frac{\alpha^4 B^2}{2m^4\mathrm{L}}\,(\mathrm{Rs}-\mathrm{Ra}\,\mathrm{L})^2 + \frac{\alpha^2}{m^2}\,(\mathrm{Rs}-\mathrm{Ra}\,\mathrm{L}) + \left(\frac{m^2}{P}+P\right)\frac{m^2}{P}\right]. \tag{4.13}$$

In order to construct the neutral curves for the complex conjugate modes (the case of oscillatory instability), we solve a system of equations (4.12) and (4.13) for the thermal Rayleigh number Ra and the frequency of neutral oscillations s. Neutral curves in the parameter space (Ra, α) obtained by the method of averaging for the cases of monotonic and oscillatory instability are presented in section 5. Note that, for the system of a pure fluid saturating a porous medium in the presence of vibration, instead of a cubic equation (4.10) we have a quadratic equation for the growth rate σ (see [14]), which can be used to prove that oscillatory instability is not possible in this case. Further analysis of (4.11) and (4.12) allows additional conclusions concerning the stability of the system under the influence of vibration and the presence of solute in the mixture. For example, under zero gravity, vertical vibration can cause instability in the binary mixture case, but not in the case of pure fluid.

5. Numerical results. The purpose of the performed numerical computations is twofold. First, we investigate the behavior of the resonant instability regions of the basic state (2.7), obtained by the method of continued fractions. Second, we find the range of vibration parameter values that provide close agreement between the results obtained by the method of continued fractions and the averaging method.

The computer code (in MATLAB) for constructing the neutral curves by the method of continued fractions solves (3.17) for the case of synchronous modes and (3.18) for the case of subharmonic modes. The continued fractions are truncated once the desired precision (10^{-4} for this paper) is achieved. For all the figures in this section, parameters not specified on the plots are chosen to be $\mathrm{Pr} = 0.733$, $l = 1$, and $c = 1$.

Neutral curves in the parameter space (Ra, α) obtained by the method of continued fractions for the synchronous and subharmonic modes are presented in Figures 5.1 and 5.2. Negative values of the thermal Rayleigh number Ra, corresponding to heating of the layer from the top, are considered in both cases. The concentration Rayleigh number Rs is chosen to be positive in Figure 5.1 (corresponding to higher concentration of the heavier component on the bottom of the layer) and negative in Figure 5.2 (corresponding to higher concentration of the heavier component on the top of the layer). Regions inside the neutral curves indicate the resonant instability of the quasi-equilibrium basic state (2.7). The alternating pattern of the instability regions

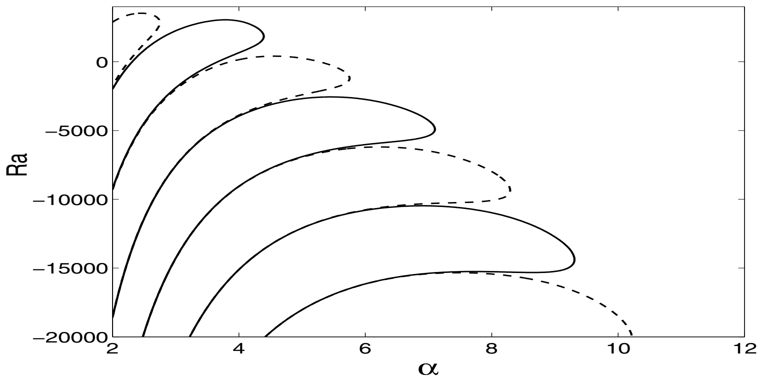


Fig. 5.1. *Neutral curves of* Ra *versus* α *for synchronous (solid line) and subharmonic (dashed line) modes, obtained by the method of continued fractions;* $\omega = 30,\ \eta = 3,\ \mathrm{Rs} = 5000,\ \mathrm{L} = 0.5.$

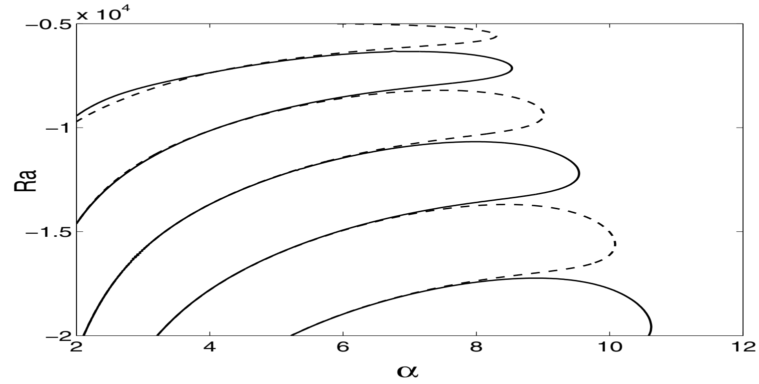


Fig. 5.2. *Neutral curves of* Ra *versus* α *for synchronous (solid line) and subharmonic (dashed line) modes, obtained by the method of continued fractions;* $\omega = 30,\ \eta = 3,\ \mathrm{Rs} = -5000,\ \mathrm{L} = 0.5.$

corresponding to synchronous and subharmonic modes can be observed in Figures 5.1 and 5.2. The values of parameters in Figures 5.1 and 5.2 correspond to the stable regime of the basic state for the case with no vibration (see [7]). Hence, the existence of the instability regions demonstrates that vertical vibration can destabilize a stable system by inducing convection in it.

Comparison of Figures 5.1 and 5.2 shows that decreasing the value of Rs (with the values of all the other parameters being fixed) causes the regions of instability to change shape (become more elongated) and move toward lower values of Ra along the vertical axis. In general, how (and if) the shape of the instability regions changes with Rs depends on the value of parameter $\mathrm{L} = \mathrm{Le}\,\varphi/\varkappa$. The change in shape observed by comparison of Figures 5.1 and 5.2 corresponds to $\mathrm{L} = 0.5$ or, more generally, $\mathrm{L} < 1$. When $\mathrm{L} > 1$, the regions of instability become more elongated as the value of Rs increases. When $\mathrm{L} = 1$, the instability regions do not change in shape with varying of Rs but simply move upward or downward along the vertical axis. The fact that the change in the instability regions depends on L can be observed by considering formula (3.12).

Neutral curves (synchronous modes) in the parameter space (Ra, α) obtained by the method of continued fractions and the averaging method are presented in Figure

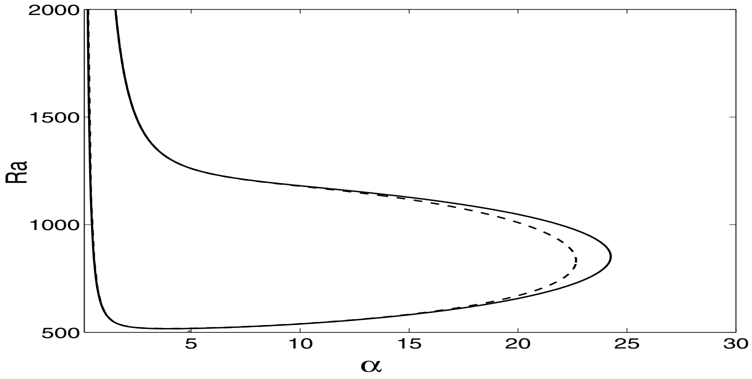


FIG. 5.3. *Neutral curves of* Ra *versus* α *for synchronous modes, obtained by the method of continued fractions (dashed line) and by the method of averaging (solid line);* $\omega = 5000$, $\eta = 500$ *for the method of continued fractions;* $\mathrm{Rs} = 500$, $\mathrm{L} = 0.5$.

5.3. The computer code (in MATLAB) solves (3.17) for constructing the neutral curves by the method of continued fractions and (4.11) for constructing the neutral curves by the method of averaging. Positive values of Ra, corresponding to heating of the layer from below, are considered in Figure 5.3. The value of the frequency ω for the curve obtained by the method of continued fractions is chosen to be 5000, which provides reasonably close agreement with the curve obtained by the averaging method in the assumption $\omega \to \infty$. This agreement improves as the value of the frequency ω in the method of continued fractions increases. Table 5.1 provides some illustrative examples using the data in Figure 5.3.

TABLE 5.1

Values of Ra *from the lower neutral curves in Figure* 5.3 *obtained by the method of continued fractions and the method of averaging (rows with* $\omega = 5000$*). Values of* Ra *obtained by setting* $\omega = 10000$ *are shown for comparison, and percent errors are shown for the method of averaging relative to the method of continued fractions.*

α	ω	Ra cont. fractions	Ra averaging method	relative error (%)
5	5000	518.54	517.78	0.15
5	10000	534.61	534.19	0.08
15	5000	582.91	581.39	0.26
15	10000	679.19	678.78	0.06

The region inside the neutral curves corresponds to instability of the basic state (2.7). The whole range of values of Ra in Figure 5.3 corresponds to the monotonic instability regime of the basic state in the unmodulated case. Hence, based on this figure, we can conclude that vertical vibration can reduce the instability region of the basic state and therefore stabilize the system.

An interesting phenomenon, which is discovered using the averaging method in the binary mixture case (and is not present in the case of a pure fluid), is the existence of closed regions of instability of the basic state (2.7). Such instability regions in the parameter space (Ra, α) are shown in Figures 5.4 and 5.5 for the case of monotonic instability and oscillatory instability, respectively. Analytically, existence of these closed regions can be predicted by considering (4.11) and (4.12) as quadratic equations for Ra. Double roots of these equations indicate the points in the parameter

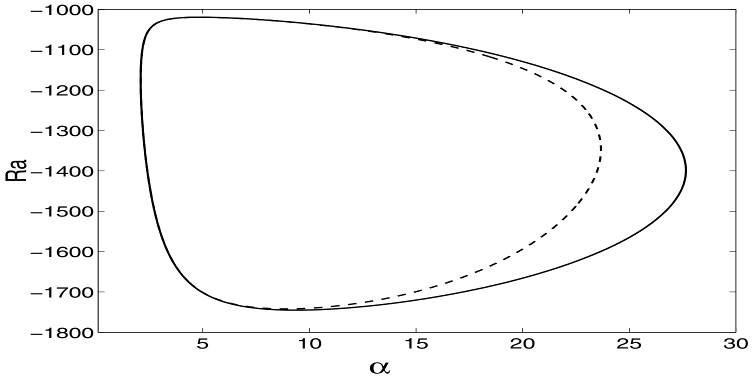


FIG. 5.4. *Neutral curves of* Ra *versus* α *for synchronous modes, obtained by the method of continued fractions (dashed line) and by the method of averaging (solid line);* $\omega = 5000,\ \eta = 500$ *for the method of continued fractions;* $\mathrm{Rs} = -1000,\ \mathrm{L} = 0.5$.

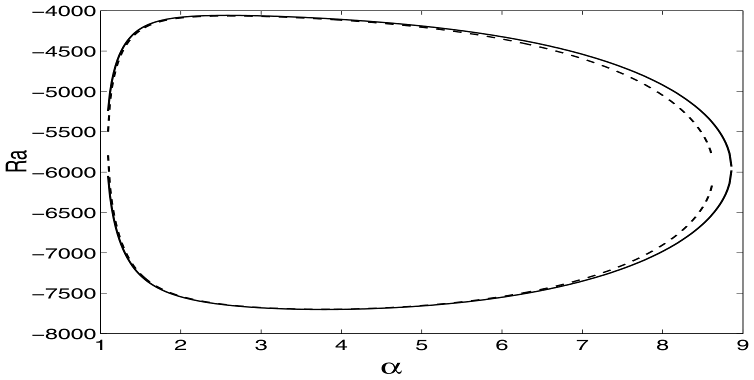


FIG. 5.5. *Neutral curves of* Ra *versus* α *for complex conjugate modes, obtained by the method of continued fractions (dashed line) and by the method of averaging (solid line);* $\omega = 2000,\ \eta = 400$ *for the method of continued fractions;* $\mathrm{Rs} = -2000,\ \mathrm{L} = 0.5$.

space where two branches of the neutral curves $\mathrm{Ra}(\alpha)$ meet, forming closed regions of instability. The fact that these regions are closed makes it difficult to locate them. To our knowledge, they were not discovered in any previous works on this problem. These regions grow in size with the growth in the absolute value of Rs, demonstrating that convective instability in the binary mixture in the presence of vibration is affected by variations in the solute concentration.

As a summary of the numerical results, sufficiently intensive vertical vibration can destabilize the system that is stable in the unmodulated case by inducing convection or stabilize an unstable system by delaying or even suppressing convection. In addition, even in the presence of vibration, the onset of convection in the system is affected by variations in the concentration of the solute in the mixture. Therefore, by varying the vibration parameters (frequency and amplitude) and the concentration of the solute, we can control convective instability in a horizontal layer of a binary fluid mixture saturating a porous medium.

Acknowledgments. I wish to thank S. Friedlander and Y. Latushkin for their helpful discussions and suggestions during preparation of this paper.

REFERENCES

[1] D. Nield, *Onset of thermohaline convection in a porous medium*, Water Resources Res., 11 (1968), pp. 553–560.

[2] J. Taunton and E. Lightfoot, *Thermohaline instability and salt fingers in a porous medium*, Phys. Fluids, 15 (1972), pp. 748–753.

[3] N. Rudraiah, P. Srimani, and R. Friedrich, *Finite amplitude convection in a two-component fluid saturated porous layer*, Int. J. Heat Mass Transfer, 25 (1982), pp. 715–722.

[4] H. Brand and V. Steinberg, *Convective instabilities in binary mixtures in a porous medium*, Phys. A, 119 (1983), pp. 327–338.

[5] B. Murray and C. Chen, *Double-diffusive convection in a porous medium*, J. Fluid Mech., 201 (1989), pp. 147–166.

[6] M. Mamou and P. Vasseur, *Thermosolutal bifurcation phenomena in porous enclosures subject to vertical temperature and concentration gradients*, J. Fluid Mech., 395 (1999), pp. 61–87.

[7] D. Nield and A. Bejan, *Convection in Porous Media*, Springer-Verlag, New York, 1998.

[8] S. Zen'kovskaya, *Effect of high-frequency vibration on filtrational convection*, Prikl. Math. Tech. Fiz., 5 (1992), pp. 83–88.

[9] S. Zen'kovskaya and T. Rogovenko, *Filtration convection in a high-frequency vibration field*, J. Appl. Mech. Tech. Phys., 40 (1999), pp. 379–385.

[10] M. Malashetty and V. Padmavathi, *Effect of gravity modulation on the onset of convection in a fluid and porous layer*, Int. J. Engrg. Sci., 35 (1997), pp. 829–840.

[11] G. Bardan and A. Mojtabi, *On the Horton–Rogers–Lapwood convective instability with vertical vibration: Onset of convection*, Phys. Fluids, 12 (2000), pp. 2723–2731.

[12] A. Jounet and G. Bardan, *Onset of thermohaline convection in a rectangular porous cavity in the presence of vertical vibration*, Phys. Fluids, 13 (2001), pp. 3234–3246.

[13] S. Govender, *Stability of convection in a gravity modulated porous layer heated from below*, Transp. Porous Media, 57 (2004), pp. 113–123.

[14] N. Strong, *Effect of vertical modulation on the onset of filtration convection*, J. Math. Fluid Mech., 10 (2008), pp. 488–502.

[15] V. Yakubovich and V. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Wiley, New York, 1975.

[16] L. Meshalkin and Y. Sinai, *Investigation of the stability of a stationary solution of a system of equations for the plane movement of an incompressible viscous liquid*, Appl. Math. Mech., 25 (1961), pp. 1140–1143.

[17] V. Yudovich, *Example of secondary stationary flow or periodic flow appearing while a laminar flow of a viscous incompressible fluid loses its stability*, Appl. Math. Mech., 29 (1965), pp. 453–467.

[18] G. Markman and V. Yudovich, *Numerical investigation of the origin of convection in the layer of fluid subjected to the external forces periodic in time*, Izv. AN USSR, MZG, 3 (1972), pp. 81–86.

[19] V. Yudovich, S. Zen'kovskaya, V. Novossiadliy, and A. Shleykel, *Parametric excitation of waves on a free boundary of a horizontal fluid layer*, C. R. Mecanique, 331 (2003), pp. 257–262.

[20] A. Khovanskii, *The Application of Continued Fractions and Their Generalizations to Problems in Approximation Theory*, P. Noordhoff, Ltd., Groningen, The Netherlands, 1963.

[21] N. Bogoliubov and Y. Mitropolsky, *Asymptotic Methods in the Theory of Non-linear Oscillations*, Hindustan Publishing Corporation, Delhi, India, 1961.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1277–1308

TIME-DOMAIN METHODS FOR DIFFUSIVE TRANSPORT IN SOFT MATTER*

JOHN FRICKS†, LINGXING YAO‡, TIMOTHY C. ELSTON§, AND
M. GREGORY FOREST¶

Abstract. Passive microrheology [T. G. Mason and D. A. Weitz, *Phys. Rev. Lett.*, 74 (1995), pp. 1250–1253] utilizes measurements of noisy, entropic fluctuations (i.e., diffusive properties) of micron-scale spheres in soft matter to infer bulk frequency-dependent loss and storage moduli. Here, we are concerned exclusively with diffusion of Brownian particles in viscoelastic media, for which the Mason–Weitz theoretical-experimental protocol is ideal and the more challenging inference of bulk viscoelastic moduli is decoupled. The diffusive theory begins with a generalized Langevin equation (GLE) with a memory drag law specified by a kernel. We start with a discrete formulation of the GLE as an autoregressive stochastic process governing microbead paths measured by particle tracking. For the inverse problem (recovery of the memory kernel from experimental data) we apply time series analysis (maximum likelihood estimators via the Kalman filter) directly to bead position data, an alternative to formulas based on mean-squared-displacement statistics in frequency space. For direct modeling, we present statistically exact GLE algorithms for individual particle paths as well as statistical correlations for displacement and velocity. Our time-domain methods rest upon a generalization of well-known results for a single-mode exponential kernel to an arbitrary M-mode exponential series, for which the GLE is transformed to a vector Ornstein–Uhlenbeck process.

Key words. generalized Langevin equation, maximum likelihood, Kalman filter, microrheology, anomalous diffusion, time series analysis

AMS subject classifications. 62M09, 62M10, 60G17, 60G20, 65C30

DOI. 10.1137/070695186

1. Introduction.

In this paper we focus on the diffusive transport of micron-scale particles in viscoelastic media. We are motivated by applications to pathogen or drug transport in pulmonary liquids (mucus) or in other biological protective barriers. We are interested in inverse methods (inference of diffusive transport properties from the primitive experimental data) and in direct simulation tools to generate both experimental time series and statistical properties such as mean-squared-displacement (MSD) and velocity autocorrelations.

To accomplish these goals, we borrow the theoretical and experimental framework from passive, single-particle microrheology as proposed by Mason and Weitz [12]. Their goal was more ambitious: from diffusive transport statistics (MSD) of dispersed microbeads, they infer bulk viscoelastic properties of the material. The Mason–Weitz (MW) theory thus combines two essential elements: a generalized Langevin equation (GLE) with a memory drag law to model the diffusion process, together with a gen-

*Received by the editors June 22, 2007; accepted for publication (in revised form) June 9, 2008; published electronically February 19, 2009. This research was supported in part by NIH R01 HL077546-01A2, NIH R01-GM078994, ARO 47089-MS-SR, and NSF grants DMS-0604891, DMS-0502266, DMS-0403040, DMS-0308687. The first and second authors contributed equally.
http://www.siam.org/journals/siap/69-5/69518.html

†Department of Statistics, Penn State University, University Park, PA 16802 (fricks@stat.psu.edu).

‡Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (yaol@math.utah.edu).

§Corresponding author. Department of Pharmacology, University of North Carolina, Chapel Hill, NC 27599-7365 (telston@amath.unc.edu).

¶Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250 (forest@amath.unc.edu).

eralized Stokes–Einstein relation (GSER) that relates the GLE memory kernel to the bulk viscoelastic modulus of the medium. We adopt only the first element, since we are exclusively interested in particle diffusion, thereby avoiding the harder problem of a direct relationship between diffusive properties and dynamic bulk moduli of the host material. The time series methods applied here are ideal for single-particle tracking experiments, which our colleagues R. Superfine, D. Hill, and J. Cribb (all at UNC) perform.

There are numerous complexities in soft matter, and especially biological materials, that frustrate a direct association of the diffusive memory kernel with the bulk viscoelastic modulus. Particle surface chemistry with the host material, particle size relative to material network lengthscales (e.g., mesh size), and heterogeneity each present nontrivial challenges. However, these issues are all circumvented for our less ambitious goal: to infer diffusive transport properties from displacement path data of microbeads. Then, one simply has to focus on inference of the memory kernel in the GLE from experimental data. We therefore choose to call the GLE memory kernel a "diffusive transport modulus," to emphasize that we are not attempting to link diffusive transport properties and bulk viscoelastic moduli.

Our inverse method applies directly to path data from particle tracking experiments, namely, position time series. This has potential advantages over ensemble averaging in frequency space, the standard approach. First, the information from individual paths is utilized, and far less data is required for parameter inversion. Second, unlike traditional microrheometry, we aim to use the results of inverse characterization to directly simulate single-particle diffusion (single paths and statistics) in biological layers. For this purpose, a time-domain representation of the memory kernel is required, which our approach yields. The MW method [11, 12] yields the unilateral Fourier transform of the imaginary part of the memory kernel, followed by application of Kramers–Kronig relations to get the real part. We refer to a very nice review article by Solomon and Lu [20] for discussions of the numerical methods associated with mapping the kernel back to the time domain.

Our second goal of direct simulations of diffusive transport processes requires forethought with respect to how one will numerically implement the modulus information gained from the inversion step. In standard inverse characterization in rheology, it is sufficient to restrict data-fitting and modulus characterization in the frequency domain. For direct simulations, we need the time-domain kernel. Thus we propose a time-domain method of inversion of the memory kernel that avoids issues with inverse transforms as discussed in [20]. Indeed, our long term goal is to couple the GLE with other dynamic processes in the biological context, e.g., pathogen diffusion in advected pulmonary liquids, or general situations where there are deterministic forces and particle-particle interactions.

Another motivation for time-domain methods is the possibility of inversion from much smaller data sets, e.g., single paths which may not be sufficient for frequency binning whereas statistical analysis of individual time series data may prove sufficient. Finally, for very small volume materials there will be constraints on the amount of sample path data that can be collected (e.g., low bead volume fractions can easily introduce colloidal effects), and a low number of sample paths may not be statistically significant for ensemble averaging. Perhaps the most compelling reason for the method proposed here is that inversion is performed directly on the physically measured data. *In this paper, we present the conceptual framework and a proof-of-principle illustration of our time-domain methods, for the Langevin and generalized Langevin models.* Particle displacement data is first generated from direct GLE simulations with a prescribed diffusive transport modulus (memory kernel); we then analyze the data with the inverse methods as though the data were path data from particle tracking experiments. A comparison of prescribed versus recovered modulus parameters is the accuracy benchmark enforced in this "methods" paper. We also compute MSD statistics directly from our formulation of the GLE and show agreement with ensemble averaging of path data.

The inverse characterization strategy introduced here is based on statistical tools developed in the field of time series analysis. These tools yield the following:

i. estimates of the viscoelastic material parameters directly from single or multiple time traces of Brownian particles,
ii. standard errors for those estimated parameters, and
iii. goodness of fit criteria.

Thus, the methods convey whether the parametrized memory kernels accurately fit the data and, in practice, how many discrete modes are needed to get a best fit. We also explore protocols for experimental sampling times and their impact on parameter inversion.

We consider an exponential (Prony) series approximation to the memory kernel, which turns out to be particularly efficient for both inversion and direct simulations. Aside from special GLE kernels, such as Rouse- and Zimm-type which are special cases of the class considered here, there is very little known about the anomalous (subdiffusive scaling on intermediate timescales) behavior of Brownian particles. We refer the reader to [17, 21] for details. For this paper, we show that our direct simulation tools recover classical Rouse and Zimm scaling properties of MSD statistics when the kernel is prescribed according to the Rouse or Zimm relaxation spectra.

The remainder of the paper is organized as follows. The standard Langevin equation for a particle diffusing in a viscous fluid is presented as a tutorial to introduce the statistical methods. In particular, we illustrate the relationship between the exact Langevin quadrature solution for particle position and autoregressive (AR) time series models. We also use the Langevin equation to introduce maximum likelihood methods for performing statistical inference of the single material parameter in the Langevin model, the fluid viscosity. Furthermore, we formulate the parameter inversion methods when only partial observations of the system are measurable (position but not velocity of Brownian particles), which is the situation in microbead rheology. Next, we show how this methodology naturally extends to multivariate AR models for GLEs with memory kernels that can be written as the sum of exponentials. The single-mode exponential kernel is presented as another tutorial example of the direct and inverse methods, since this example can also be analyzed in explicit closed form. Next, 4-mode kernels, of classical Rouse and Zimm form, are used as a nontrivial illustration of the direct and inverse methods, and finally a 22-mode Rouse kernel is presented to show that the direct simulations are not limited to a sparse discrete spectrum.

A significant by-product of these investigations arises from two critical observations:

- GLEs with arbitrary finite-mode, exponential kernels are exactly integrable with a quadrature solution [7]; and
- the quadrature formula extends from the continuous GLE process to a discretized dynamics.

These two observations yield a statistically exact, discrete-time AR process model of a Brownian particle in a viscoelastic medium. The first-order Taylor approximation of this discrete process corresponds to a first-order Euler numerical integration scheme. *This class of discrete GLE models thereby provides a highly efficient and accurate direct time-domain simulation method.* We can generate realizations of Brownian particles in a viscoelastic fluid, based on matrix function evaluation rather than a low-order numerical integration of the stochastic GLE model. Furthermore, *average properties (MSD and velocity correlations) also have explicit quadrature representations, so that statistical correlations may be simulated directly, avoiding the arduous alternative of generating sample paths and then averaging.* In examples presented below, we benchmark the numerical tools by confirming agreement between the two ways of computing MSD statistics. These direct simulation results thus afford the ability to simulate time-domain experimental data of individual particles as well as statistical scaling properties of Brownian particles for any given exponential series form of the memory kernel in the GLE model.

For arbitrary M-mode kernels with $M > 1$, there is one numerical analysis result required to assure accurate computation of matrix exponentials in the discrete and continuous quadrature formulas, which we provide in the appendix. With this result, numerical simulations are carried out in the body through various explicit examples. It is worth emphasizing that *this approach*—replacing stochastic numerical integration by matrix function evaluation in a discrete GLE process, for individual paths as well as for average properties of the process—*is guaranteed to be statistically correct, even for sufficiently long time series.* This strategy removes two dominant sources of numerical error in the direct problem of time-domain simulation: the error at each time step from a low-order integration method instead of an exponential-order method, and the cumulative error in time-stepping, which is completely avoided. Because many generic memory kernels can be approximated to arbitrary accuracy with a sum of exponentials, this simulation method should find utility in diverse applications outside of pulmonary liquids. The range of diffusive dynamic scaling behavior of individual Brownian particle paths, and of ensemble averages, is a topic for future study to understand the range of diffusive transport statistics possible for GLEs with exponential series kernels. The known theoretical results for Rouse and Zimm spectra will be illustrated and confirmed below as rigorous benchmarks on our direct simulation strategy, as well as for inverse characterization benchmarks of the maximum likelihood method.

2. The Langevin equation and statistical methods.

In this section, we review the basic properties of the classical Langevin equation for a microscopic particle diffusing in a viscous fluid, as a transparent context for introducing our statistical approach. The solution of the Langevin equation can be exactly represented as a Gaussian AR statistical model (cf. [8]). Thus, a maximum likelihood approach can be used to estimate model parameters from time series data. To illustrate the methodology, the statistical tools are developed first, assuming that the velocity of the particle is directly measured. However, in microscopy experiments the particle position (and not velocity) is measured. Thus, using standard techniques, we next generalize the statistical framework to a two-dimensional Langevin equation for both position and velocity, *in which only position observations are required for statistical inference of model parameters.* All advantages of maximum likelihood estimation are preserved in this formulation, which we illustrate numerically.

2.1. The Langevin equation and quadrature solution.

The scalar Langevin equation for a diffusing particle with velocity v is

$$m\frac{dv}{dt} = -\xi v + \sqrt{2k_BT\xi}f(t), \tag{2.1}$$

where m is the particle mass, k_BT is the Boltzmann constant times the absolute temperature, and the friction coefficient $\xi = 6\pi a\eta$, where a is the radius of the particle and η is the viscosity of the fluid. The stochastic term $f(t)$ is taken to be Gaussian white noise with zero mean and covariance,

$$\langle f(t)f(s)\rangle = \delta(t-s). \tag{2.2}$$

Equation (2.1) represents a 2-parameter linear stochastic differential equation (SDE), written equivalently in the standard form of an Ornstein–Uhlenbeck process:

$$\frac{dv(t)}{dt} = -\alpha v(t) + \sigma f(t), \tag{2.3}$$

where the two parameters in the process are

$$(\alpha, \sigma) = (\xi/m, \sqrt{2k_BT\xi/m^2}). \tag{2.4}$$

Ornstein–Uhlenbeck processes have several important properties—Markovian, stationary (given an appropriate initial condition), and Gaussian—that are amenable to mathematical and statistical analysis.

- If the initial velocity $v(0)$ is normally distributed with mean zero and variance $\sigma^2/(2\alpha)$,

  $$v(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\alpha}\right), \tag{2.5}$$

  then $v(t)$ has the same distribution for all t, and the velocity autocorrelation function (ACF) is given by

  $$\langle v(t)v(s)\rangle = \frac{\sigma^2}{2\alpha}e^{-\alpha|t-s|}. \tag{2.6}$$

- An Ornstein–Uhlenbeck process can be written in terms of a stochastic integral:

  $$v(t) = e^{-\alpha t}v(0) + \sigma\int_0^t e^{-\alpha(t-s)}f(s)ds, \tag{2.7}$$

  which is a quadrature solution to the SDE (2.3).
- This representation is useful, as shown below, for developing efficient statistical techniques for estimating the parameters α and σ from time series data sampled on finite intervals.
- From the exact solution, the tracer position $x(t)$ is given by

  $$x(t) = x_0 + \int_0^t v(s)ds, \tag{2.8}$$

  where $x_0 = x(t=0)$. The variance of the tracer position (MSD) is likewise explicit [3]:

  $$\langle[x(t) - x(0)]^2\rangle = \frac{2k_BT}{\alpha m}\left[t - \frac{1}{\alpha}(1 - e^{-\alpha t})\right]. \tag{2.9}$$

Next we introduce and apply statistical methods that take advantage of the Gaussian evolution and integrability of the Langevin equation to recover α and σ from time series data. These features will be shown in subsequent sections to carry over to the GLE and thereby to inversion of viscoelastic parameters from tracer time series data.

2.2. AR processes and exact discrete Langevin equations. Suppose we want to match Brownian tracer experimental data with a discrete model of the Langevin equation (2.3), where the discrete time step Δ has to be sufficiently small to resolve the underlying stochastic process. The velocity of a particle diffusing in a viscous fluid can be modeled by discretizing equation (2.3) using an Euler approximation, which yields

$$v_n - v_{n-1} \approx -\alpha v_{n-1}\Delta + \sigma\sqrt{\Delta}\epsilon_n, \tag{2.10}$$

where ϵ_n is a sequence of independent standard normal random variables and $v_n = v(n\Delta)$. Rearranging the above equation yields

$$v_n \approx (1-\alpha\Delta)v_{n-1} + \sigma\sqrt{\Delta}\epsilon_n. \tag{2.11}$$

With this discretization, v_n is a *first-order autoregressive (AR) process*. An AR process is one in which the current observation is a weighted sum of the previous observations plus a noise term that is independent of previous noise terms. Alternatively, we can *exploit the quadrature solution* (2.7) and replace the approximate discretization by the *exact discrete Langevin process*,

$$v_n = e^{-\alpha\Delta}v_{n-1} + \epsilon_n, \tag{2.12}$$

where ϵ_n, $n = 1, \ldots, N$, is a sequence of independent standard Gaussian random variables with variance

$$s(\alpha,\sigma) = \sigma^2 \frac{1-e^{-2\alpha\Delta}}{2\alpha}. \tag{2.13}$$

The Euler approximation is recovered as a first-order Taylor series expansion of the coefficients in this exact discretization. The advantages of this exact discretization are that one can accurately generate sample paths, and furthermore, the time series are guaranteed to be statistically consistent with the process (which might otherwise be polluted by cumulative errors in a numerical integration scheme). We will apply this discrete process to simulate an experiment, from which experimental time series are extracted by sampling the full data set.

2.3. Maximum likelihood methods for parameter inversion. We turn now to maximum likelihood methods which give a general framework for obtaining point estimators and standard errors for the model parameters, α and σ, given a time series $v_0, v_1, \ldots, v_N$. The likelihood function is computed from the joint probability density for an observed velocity time series. Noting that the time series is Markov, that the conditional distribution of v_n given v_{n-1} is normal with mean $e^{-\alpha\Delta}v_{n-1}$ and variance (2.13), and assuming that the initial velocity v_0 is known, the likelihood function is given by

$$\begin{aligned} L(\alpha,\sigma) &= g(v_1,\ldots,v_N|v_0,\alpha,\sigma) \\ &= \prod_{n=1}^{N} h(v_n|v_{n-1},v_0,\alpha,\sigma) \\ &= (2\pi s(\alpha,\sigma))^{-n/2}\exp\left(-\sum_{n=1}^{N}\left(\frac{v_n - e^{-\alpha\Delta}v_{n-1}}{2s(\alpha,\sigma)}\right)^2\right), \end{aligned}$$

where $g(\cdot|v_0, \alpha, \sigma)$ is the joint density of $v_1, \dots, v_N$ and $h(\cdot|\cdot, v_0, \alpha, \sigma)$ is the transition density for the process. Given a sequence of velocity measurements, the likelihood function is numerically maximized to obtain estimates, $\hat{\alpha}$ and $\hat{\sigma}$, for α and σ. Hereafter in the paper, parameter estimates are denoted by $\hat{\cdot}$.

One of the benefits of maximum likelihood estimation is that, under fairly general conditions to be given in the appendix, asymptotic probability distributions for these estimators may be obtained. Note that while α is not random, $\hat{\alpha}$ depends on the random time series $v_0, \dots, v_N$ and is a random variable; given a new time series, one obtains a new realization of the random variable. In the present context, we know a priori that the estimator $\hat{\alpha}$ is asymptotically (for long time series, i.e., large number of observations N) normal with mean equal to the true parameter α and variance of $\hat{\alpha}$ equal to $1/N(-\partial_\alpha^2 \log L(\alpha, \sigma))^{-1}$. We obtain an estimate for the variance of $\hat{\alpha}$ by numerically calculating the derivative of the log likelihood function at the maximized value.

We emphasize that model parameters may be estimated from a single time series of the process; this will be illustrated in the proof-of-principle illustrations below. If that single-particle path is sufficiently long, then the MW approach and our approach should be consistent. (A final example addresses this point.) If multiple paths are available and they are presumed to be independent, the overall likelihood function will be defined as the product of likelihood functions for the individual paths, and maximum likelihood estimators may be obtained as before including the additional observations. This methodology will be valid assuming statistical independence of the paths. The methods introduced here can be applied even if the data set is not large; this corresponds either to a large Δ or a low number of iterations in the discrete process. We will return to this issue below in a discussion of over- and underresolution of the underlying stochastic process, and in comparisons of quality of fits versus number of observations.

2.4. Extension to the full system of position and velocity. In general, microrheology experiments measure the position of the particle, not the velocity. It is of course unwise to approximate the velocity by differencing the experimental data; information is lost and unnecessary errors are introduced. Alternatively, we formulate a vector Langevin model for the position and velocity of the particle, and then develop maximum likelihood methods assuming only partial observations of the process variables. Specifically, we can observe $x_0, x_1, \dots, x_n$ but cannot observe $v_0, v_1, \dots, v_n$. The system can be written in vector form as

$$\frac{d}{dt}Y(t) = AY + Kf(t), \tag{2.14}$$

where

$$Y = \begin{pmatrix} x(t) \\ v(t) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ 0 & -\alpha \end{pmatrix}, \quad K = \begin{pmatrix} 0 \\ \sigma \end{pmatrix}, \tag{2.15}$$

and $f(t)$ is a scalar Gaussian white noise process defined above. The *quadrature solution* to (2.14) is [15]

$$Y(t) = e^{At}Y(0) + \int_0^t e^{A(t-s)}Kf(s)ds. \tag{2.16}$$

As noted above, special properties of the exact solution can be exploited when performing parameter estimation. The process is Gaussian and therefore uniquely defined by its mean and covariance. So, given an initial condition $Y_0 = Y(0)$ and a time increment Δ, we can determine the exact distribution of $Y_1 = Y(\Delta)$ and by iteration define a vector AR process, as in (2.12) above.

Conditioning on Y_{n-1}, the distribution of Y_n is Gaussian with mean $e^{A\Delta}Y_{n-1}$ and covariance matrix [8, 15]

$$S(\Delta) = \int_0^\Delta e^{A(\Delta-s)}KK^T e^{A^T(\Delta-s)}ds. \tag{2.17}$$

Furthermore, it is straightforward to generate exact realizations of the stochastic process at finite time intervals, with the caveat that one must be able to accurately calculate S. (For A, K in (2.15), this is trivial; for the GLE of viscoelastic fluids, we address this issue in section 3.1.) For a particle starting in state Y_0, we generate a Gaussian vector ϵ_n with covariance matrix S and add this to $e^{A\Delta}Y_0$ to obtain Y_1, and then simply iterate this procedure. That is,

$$Y_n = e^{A\Delta}Y_{n-1} + \epsilon_n, \tag{2.18}$$

where ϵ_n is an independent sequence of zero mean Gaussian random vectors with covariance S. Thus, we have an AR representation for the vector process $Y_0, \ldots, Y_N$ associated with the scalar process (2.12).

2.5. The likelihood function for position measurements. Now that we have cast the Langevin model in the form of a vector AR process, we are in position to calculate the appropriate likelihood function for estimating parameters, given a time series of particle positions $x_0, x_1, \ldots, x_N$. In this section, we outline key steps in the derivation of the likelihood function, leaving a detailed derivation for the appendix. The derivation relies on the Kalman filter, which was developed to estimate the current state of a dynamical system from noisy time series data of partial observations of the process. (This use of the Kalman filter as a method to calculate the likelihood function has become standard, and further discussion can be found in [2] and [8].) Recall that discrete observations generated from the Langevin equation satisfy (2.18), where the noise has a covariance structure given by (2.17). Experimentally, only the position of the particle is observed, and no other components of the vector Y. That is, at the nth time interval the observable is

$$x_n = CY_n, \quad C = \begin{pmatrix} 1 & 0 \end{pmatrix}. \tag{2.19}$$

Assuming that the model parameters, Θ, are known, a Kalman filter is generally used to recursively estimate the current state, Y_n, given the observations $x_1, \ldots, x_n$. Using this and the AR structure of the process, we may also give a predictive density for Y_{n+1} given $x_1, \ldots, x_n$. From this we may obtain the density of x_{n+1} given $x_1, \ldots, x_n$, which we denote by $h(x_{n+1}|x_m, m < n+1, \Theta, x_0)$. We may then decompose the joint density for the time series into a product of these conditional densities and obtain

$$g(x_1, x_2, \ldots, x_N|\Theta, x_0) = \prod_{n=2}^{N} h(x_n|x_m, m < n, \Theta, x_0). \tag{2.20}$$

Because the process is Gaussian, the above equation can be rewritten as

$$\begin{aligned} -\log L(\Theta) &= -\log g(x_1, x_2, \ldots, x_N|\Theta, x_0) \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(\log 2\pi + \log Q_{n-1} + \frac{(x_n - \hat{x}_{n|n-1})^2}{Q_{n-1}}\right), \end{aligned} \tag{2.21}$$

where the conditional mean and variance of x_n given $x_1, \dots, x_{n-1}$ are

$$\hat{x}_{n|n-1} = Ce^{A\Delta}\hat{Y}_{n-1} \tag{2.22}$$

and

$$Q_{n-1} = CR_{n-1}C^t, \tag{2.23}$$

respectively, and the matrix R_n is defined in the appendix. Therefore, once we have $x_0, x_1, \dots, x_N$ we may numerically maximize this likelihood function with respect to the parameters to obtain an estimate for Θ. An important feature of this Kalman derivation of the likelihood function is that it may be calculated recursively; this dramatically reduces the time necessary to calculate the likelihood function since we do not have to calculate the full covariance matrix of the entire time series. Use of the Kalman filter to calculate the likelihood function of dependent data is a common procedure in time series analysis and is the most accurate and efficient method to calculate the likelihood function for a number of common models such as the ARIMA model [6, 18].

This method requires numerical calculation of the matrices S and $e^{A\Delta}$, but this calculation has to be done only *once* for each trial parameter set in the maximization process. This numerical calculation is, of course, trivial for 2×2 systems, but presents a potential limitation for the GLE, which we will soon formulate in this precise vector AR setting, and where the matrix size scales with the number of exponential modes. Below, we overcome this potential limitation due to the special form of the matrices that arise for GLEs with exponential kernels.

As with the univariate case, there are asymptotic results for the distribution of our maximum likelihood estimators $\hat{\Theta}$. Under certain reasonable conditions given in the appendix, $\hat{\Theta}$ is asymptotically normal with mean Θ and covariance given by $\text{cov}(\hat{\Theta}) = 1/N(-\nabla \log L(\Theta))^{-1}$, which may be approximated by numerical evaluation of the quantity $1/N(-\nabla^2 \log L(\hat{\Theta}))^{-1}$. Thus, to build a $1-\alpha$ confidence interval for θ_m, we start with

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\Theta}_m - \theta_m}{\text{cov}(\hat{\Theta})_{m,m}} \leq z_{\alpha/2}\right) \approx 1 - \alpha, \tag{2.24}$$

where $z_{\alpha/2}$ is the value that satisfies $P(Z > z_{\alpha/2}) = \alpha/2$ and Z is a standard Gaussian random variable. We use the notation $A_{m,n}$ to denote the element in the mth row and nth column of the matrix A. Some algebra yields

$$\theta_m \in (\hat{\Theta}_m - z_{\alpha/2}\text{cov}(\hat{\Theta})_{m,m}, \hat{\Theta}_m + z_{\alpha/2}\text{cov}(\hat{\Theta})_{m,m}), \tag{2.25}$$

which is the desired confidence interval for θ_m.

2.6. The autocorrelation function (ACF). A common diagnostic tool for determining important timescales in time series data is the *discrete autocorrelation function.* This function represents a scaled and discretized estimate of the true autocovariance function

$$\text{Cov}\,(U(t)U(s)) = \langle U(t)U(s)\rangle - \langle U(t)\rangle\langle U(s)\rangle. \tag{2.26}$$

For a discrete time series $U_1, \dots, U_N$, where $U_k = U(k\Delta)$ and the data is normalized to have mean zero, the discrete ACF is defined to be

$$ACF(j) = \frac{\sum_{n=j+1}^{N} U_n U_{n-j}}{\sum_{n=1}^{N} U_n^2}.$$

From now on the acronym ACF denotes the discrete autocorrelation function unless explicitly stated otherwise. Note that for zero time lag, the ACF is normalized to one. A general guide for verifying that a process is white noise (independent identically distributed sequence of random variables) is that for all lags greater than or equal to one the ACF will be less than $2/\sqrt{N}$, where N is the number of observations [19]. We illustrate the application of the ACF diagnostic in examples below.

2.7. Illustration of the statistical toolkit. We present a simple example, Brownian diffusion and simple Langevin dynamics, to show how these methods work and test their accuracy. The example illustrates the importance of the experimental sampling time relative to the physical timescales in the model. We always assume (and enforce in numerical simulations) that the discrete time step Δ in the direct simulation of sample paths is small enough to resolve the stochastic fluctuation timescales in the model. This yields a faithful resolution of the physical process from which we can then sample the resolved data on any coarse timescale, analogous to an experimental sampling time. With these protocols, we are able to provide measures and indicators of experimental over- and undersampling.

Throughout the paper, we measure time in milliseconds (ms), mass in milligrams (mg), and length in microns (μm). Consider a neutrally buoyant particle of diameter 1 μm and mass 5×10^{-10}mg moving in a fluid with viscosity 1.5 Pa-s (similar to glycerol). This corresponds to $\alpha = 26 \times 10^6 (\text{ms})^{-1}$ and $\sigma = 65 (\text{ms})^{-3/2}$. First, we simulate the exact discrete Langevin process (2.17), (2.18) for a highly resolved time step $\Delta = 10^{-10}$ms, which is three orders of magnitude smaller than the viscous timescale set by the drag coefficient, $\alpha^{-1} = \text{m}/\zeta \approx 0.37 \times 10^{-7}$ms. We generate one sample path with 10^5 data points. The examples to follow will strobe this data set at the prescribed lag Δ; if Δ is $10^{-10+\delta}$, then each observation corresponds to 10^{δ} numerical time steps.

The ACF is first computed using a coarse sampling time $\Delta = 5 \times 10^{-7}$ms, which is 13.4 times the viscous timescale α^{-1}. The process yields the ACF signature of white noise, Figure 1(top). That is, the ACF nearly approximates a delta distribution versus lag with most of the weight at zero lag time, and therefore at this sampling interval the process appears to be white noise. On the other hand, if the sampling interval is shortened ($\Delta = 10^{-8}$ms) so that it is consistent with the viscous timescale, then the ACF falls off exponentially, as in Figure 1(bottom).

Next, we use maximum likelihood methods to generate the estimators $\hat{\alpha}$ and $\hat{\sigma}$ for five decades of lags Δ (Figure 2). Note that the estimator (open circles) is most accurate and the variance (vertical bars) is minimized when the lag time $\Delta \approx 10^{-8}$–10^{-9}ms, consistent with the ACF diagnostic (Figure 1(bottom)) showing exponential decay. Note further that the estimator $\hat{\alpha}$ degrades as Δ increases, and the variance grows, consistent with the ACF of Figure 1(top) for coarse sampling. For Δ very small, e.g., $\Delta = 10^{-10}$ms, the variance of α again grows, but the estimator remains quite accurate.

This simple example illustrates a method for choosing an appropriate time interval for sampling. If the observations are too far apart (underresolved), e.g., $\Delta = 10^{-7}$ms, then the autocovariance of the velocity is near zero after one time step. Indeed, one can compute the AR matrix

$$e^{A\Delta} = \begin{pmatrix} 1 & \frac{1-e^{-\alpha\Delta}}{\alpha} \\ 0 & e^{-\alpha\Delta} \end{pmatrix} \overset{\Delta=10^{-7}}{\approx} \begin{pmatrix} 1 & 3.7\cdot 10^{-8} \\ 0 & 1.5\cdot 10^{-6} \end{pmatrix}. \tag{2.27}$$

Looking at the discrete process (2.18) and (2.17), there is little information carried

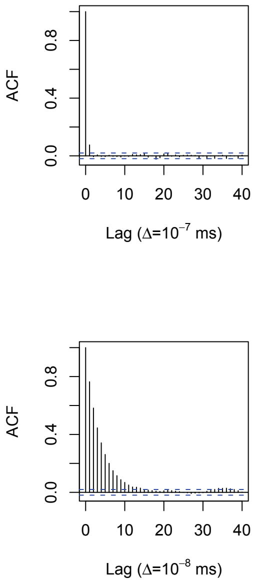


Fig. 1. *ACF of Langevin equation velocity time series: The ACF of the velocity at two different sampling intervals, one showing underresolution and the other indicating accurate resolution.*

over except the previous position, so the process is nearly a discrete white noise process. Nonetheless, the time series approaches can often still give reasonable estimates of the parameters, as shown in Figure 2. By contrast, a reasonable sampling time, like $\Delta \sim 10^{-8}$ms, will reflect an exponential ACF, signalling good resolution of the process. In the extremely improbable situation where observations are too frequent (overresolved), e.g., $\Delta = 10^{-10}$ms, then the AR matrix will be close to the identity,

$$e^{A\Delta} \overset{\Delta=10^{-10}}{\approx} \begin{pmatrix} 1 & 9.9{\cdot}10^{-5} \\ 0 & 9.9{\cdot}10^{-1} \end{pmatrix},$$

and the velocity will appear to be nonstationary with a linear decay in the ACF. These signatures of the ACF are tools that can be used with experimental data to identify an appropriate sampling time, and even to estimate the smallest physical timescale in the underlying process.

3. The GLE and statistical methods.

3.1. Mathematical framework: Quadrature solution for exponential series kernels. The starting point for modeling the diffusive properties of microscopic Brownian particles in viscoelastic materials is the GLE [1, 7, 12, 16, 22, 23]:

$$m\frac{dV(t)}{dt} = -\int_0^t \varphi(t-\tau)V(\tau)d\tau + \tilde{F}(t). \tag{3.1}$$

For passive microrheology, $\tilde{F}(t)$ is a Gaussian colored noise, correlated with the memory kernel $\varphi(t)$ through the fluctuation-dissipation relation,

$$\langle \tilde{F}(t), \tilde{F}(s)\rangle = k_B T\varphi(t-s), \quad t > s. \tag{3.2}$$

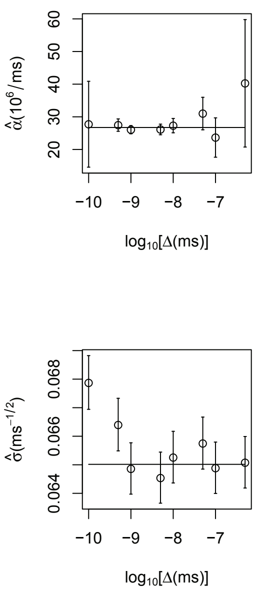


FIG. 2. *Parameter estimates versus sampling time Δ of the drag α and noise σ for the Langevin model. The bands represent* 95% *confidence intervals for the estimates. The true parameter is represented by a horizontal line.*

For consistency with the Langevin illustration, we divide both sides of (3.1) by m and redefine the memory kernel appropriately to obtain

$$\frac{dV(t)}{dt} = -\int_0^t \xi(t-\tau)V(\tau)d\tau + \sqrt{\frac{k_BT}{m}}F(t), \tag{3.3}$$

with

$$\langle F(t), F(s)\rangle = \xi(t-s), \quad t > s. \tag{3.4}$$

Throughout the remainder of the paper when we refer to the memory kernel, we will mean $\xi(\cdot)$, which is scaled by $1/m$.

In this section, we show that for a certain class of memory kernels, specifically a sum of exponentials, the GLE can be expressed as a set of coupled linear SDEs *of the same form as* (2.14), in which the velocity and position are the first two components. Therefore, all Langevin equation properties and techniques carry over immediately to the GLE. In particular, we can (1) apply maximum likelihood methods for parameter estimation, (2) exactly simulate the stochastic process instead of low-order numerical integration, and (3) write down explicit formulas for statistical quantities of interest, such as ACFs for position and velocity.

Suppose the memory kernel is a single exponential,

$$\xi(t) = ce^{-\frac{t}{\lambda}}, \quad c = \frac{6\pi aG}{m}, \tag{3.5}$$

where a and m are the particle radius and mass, which corresponds to a single-mode Maxwell fluid with shear modulus G, relaxation time λ, and zero strain rate viscosity

$\eta = \lambda G$. The noise $F(t)$, (3.3)–(3.4), for the single exponential kernel can be expressed as an Ornstein–Uhlenbeck process,

$$\frac{dF(t)}{dt} = -\frac{1}{\lambda}F(t) + \sqrt{\frac{2c}{\lambda}}f(t), \tag{3.6}$$

where $f(t)$ is white noise. Note that the Langevin equation for viscous diffusion is obtained in the limit $\lambda \to 0$; that is, (3.6) becomes (with $\xi = 6\pi a\eta$)

$$F(t) = \sqrt{\frac{2\xi}{m}}f(t). \tag{3.7}$$

Analogous to the scalar Ornstein–Uhlenbeck process (2.3), the system (3.3)–(3.6) may be solved explicitly, which has been noted in several classical references [1, 7, 16, 22, 23]. To see this, define the variable $Z(t)$ by

$$Z(t) = \int_0^t e^{-\frac{t-\tau}{\lambda}}V(\tau)d\tau, \tag{3.8}$$

which yields

$$\frac{dZ(t)}{dt} = -\frac{1}{\lambda}Z(t) + V(t). \tag{3.9}$$

Now, the full system can be written in matrix form as

$$\frac{d}{dt}Y(t) = AY(t) + KW(t) \tag{3.10a}$$

with

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -c & \sqrt{\frac{k_BT}{m}} \\ 0 & 1 & -\frac{1}{\lambda} & 0 \\ 0 & 0 & 0 & -\frac{1}{\lambda} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\frac{2c}{\lambda}} \end{pmatrix}, \tag{3.10b}$$

$$Y(t) = (X(t), V(t), Z(t), F(t))^T, \tag{3.10c}$$

and $W(t)$ is a vector of independent white noise processes.

This system (3.10a)–(3.10c) is *identical* in form to (2.14), and therefore another vector Langevin equation, whose quadrature solution is given by (2.16) and (2.17) with these Y, A, and K. Following the Langevin example above, we can now generate the corresponding viscoelastic AR process for a Brownian particle with this specified memory kernel, starting from $Y_0 = Y(0)$.

More generally, suppose that the memory kernel $\xi^M(t)$ is given by an M-mode exponential series:

$$\xi^M(t) = c_1e^{-\frac{t}{\lambda_1}} + c_2e^{-\frac{t}{\lambda_2}} + \cdots + c_Me^{-\frac{t}{\lambda_M}}, \tag{3.11}$$

where $c_i = 6\pi aG_i/m$. Similarly, the total noise $F^M(t)$ can be written as

$$F^M(t) = F_1(t) + F_2(t) + \cdots + F_M(t), \tag{3.12}$$

where each $F_i(t)$ is an independent Ornstein–Uhlenbeck process characterized by the ith relaxation time λ_i. That is,

$$\frac{dF_i(t)}{dt} = -\frac{1}{\lambda_i}F_i(t) + \sqrt{\frac{2c_i}{\lambda_i}}f_i(t), \tag{3.13}$$

where $f_i(t)$, $i = 1, \dots, M$, are independent white noise processes.

Therefore, $F^M(t)$ is a mean-zero Gaussian process with covariance consistent with the fluctuation-dissipation theorem,

$$\langle F^M(t)F^M(s)\rangle = c_1 e^{-\frac{t-s}{\lambda_1}} + c_2 e^{-\frac{t-s}{\lambda_2}} + \cdots + c_M e^{-\frac{t-s}{\lambda_M}}. \tag{3.14}$$

This formulation of the GLE yields once again a vector Langevin process of the form (36), with the following definitions for Y, A, and K:

(3.15)

$$Y = \begin{pmatrix} X(t) \\ V(t) \\ Z_1(t) \\ \dots \\ Z_M(t) \\ F_1(t) \\ \dots \\ F_M(t) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & -c_1 & \dots & -c_M & \sqrt{\frac{k_BT}{m}} & \dots & \sqrt{\frac{k_BT}{m}} \\ 0 & 1 & -1/\lambda_1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & \dots & -1/\lambda_M & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & -1/\lambda_1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & -1/\lambda_M \end{pmatrix},$$

$$K = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \sqrt{\frac{2c_1}{\lambda_1}} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \sqrt{\frac{2c_M}{\lambda_M}} \end{pmatrix}.$$

Again, an *exact solution of this system* is given in the form (2.16) and (2.17) with these matrix formulas. Thus, all properties of the Langevin equation have been extended to the GLE for the class of M-mode exponential series kernels. Likewise, the machinery from section 2 applies for generating direct realizations of GLE processes and performing statistical analysis of time series for partial observations (of position).

These formulas are valuable to the extent that we can numerically calculate the matrix exponential e^A. The special form of A, (3.15), lends itself to an explicit and straightforward determination of the eigenvalues and eigenvectors for any mode number M. Furthermore, this calculation only has to be done once, both to generate the direct process (or statistics of the process) and to perform parameter inversion for each M-mode model. The procedures for computing the spectrum and then the covariance matrix are given in the appendix.

3.2. GLE direct and inverse illustration with a single exponential kernel.

We first illustrate the GLE direct and inverse strategy, analogous to the Langevin illustration in section 2, for the simplest case: a single-mode exponential kernel (3.5)

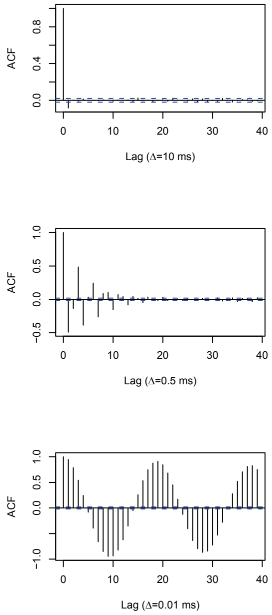


FIG. 3. *ACF versus sampling interval Δ for a GLE with single-mode exponential kernel with relaxation timescale $\lambda_1 \sim 1.5ms$. Top: Underresolved with $\Delta \sim 6\lambda_1$. Middle: Resolved with $\Delta \sim .3\lambda_1$. Bottom: Overresolved with $\Delta \sim .01\lambda_1$.*

for which the GLE is given by (3.10a)–(3.10c). We select physical parameter values as follows: $\lambda_1 = 1.546$ ms, $G_1 = 1.035 \times 10^{-5}mg/ms^2\mu$m. The model parameter c_1 then has the value $c_1 = 4.440 \times 10^{-3}$ms^{-2}. Data are generated by a direct simulation with time step Δ ms; we explore various sampling intervals relative to λ to identify signatures of over-, under-, and "good" sampling times in the ACF and the estimators $(\hat{\lambda}_1, \hat{c}_1)$. For each Δ, we generate a single sample path consisting of 5×10^4 observations, or a total experimental simulation of $5 \times 10^4 \Delta$ ms.

We begin with the effect of sampling interval Δ on the ACF for velocity, as shown in Figure 3. The data for bead velocity were created by differencing the position data for a sample path of length 50,000. The first plot in Figure 3 corresponds to a very long sampling interval (6 times the relaxation time λ_1) and shows that the velocities at consecutive time steps are nearly independent of one another. We can see this by analyzing the matrix $e^{A\Delta}$, and we notice

$$v_{n+1} \approx 0.036 v_n + \epsilon, \tag{3.16}$$

where ϵ is white noise, which explains why the ACF of velocity approximates white noise. The second plot shows a more reasonable ACF at a sampling interval $\Delta = 0.5$ms. The last ACF plot in Figure 3 corresponds to a very fast sampling interval $\Delta = 0.01$ms. Note that for this sampling rate, the ACF appears to fall off linearly, rather than exponentially as expected, indicative of a process that has been oversampled. This behavior is similar to the Langevin equation, where very short time steps yield a strong dependence from one velocity to the next. Recall that this scenario yields a likelihood function that is relatively insensitive to parameter values.

Figure 4 shows the maximum likelihood estimate $\hat{\lambda}_1$ of a single relaxation time, λ_1, from numerically generated data and demonstrates the effect of the sampling

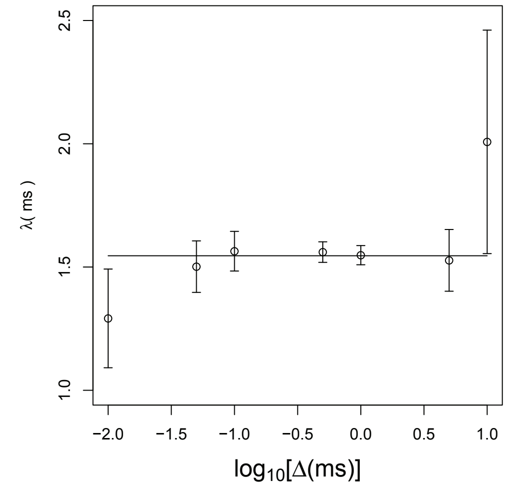


FIG. 4. *Estimators of the relaxation time* $\hat{\lambda}_1$ *versus sampling resolution* Δ*, with data taken from a direct discrete GLE simulation with a single-mode exponential memory kernel. The exact value* $\lambda_1 = 1.546ms$ *is denoted by the horizontal line. The hollow circle indicates the value of the estimator, and the error bars indicate* 95% *confidence intervals.*

interval on the estimation of λ_1. The horizontal line represents the true value of λ_1, while the error bars represent 95% confidence intervals which are symmetric about the estimate represented by open circles. As with the ordinary Langevin case, there is an optimal sampling interval. Note that the natural timescale for this parameter is on the order of milliseconds; this is approximately the sampling interval at which the minimum variance of the estimator is obtained.

It is important to note here that for each sampling rate the number of discrete observations used for inference is being held constant. This implies that the real time interval over which the observations are being taken is much shorter for the faster sampling rates and considerably longer for the slowest sampling rates. This shorter real time interval could partially explain the large variance of the estimator at these faster rates. However, one should also note that the observations taken at longer than optimal sampling intervals occur over a longer real time interval and yet also perform poorly. This demonstrates that both sampling rate and number of observations play a role in the performance of the method, which is worthy of further investigation.

In Figure 5, the estimate $\hat{c}_1$ of the model parameter c_1 versus sampling interval Δ is illustrated. As seen when estimating λ_1, the estimates improve as the sampling interval becomes longer. However, beyond the interval of Δ values in this plot the quality of the estimator declines quickly. Note that this parameter has a natural timescale of $1/\sqrt{c_1}$ which is approximately $10^{-\frac{3}{2}}$ms. Note also that there is little overlap between the very good estimates of c_1 and the good estimates of λ_1. This points to a general problem for a system with different relevant timescales. The quality of relative estimates within a parameter set will be partially determined by the sampling interval.

In Figure 6, we show the effect of the number of experimental observations on parameter estimation. Parameter estimates improve with the length of the time series

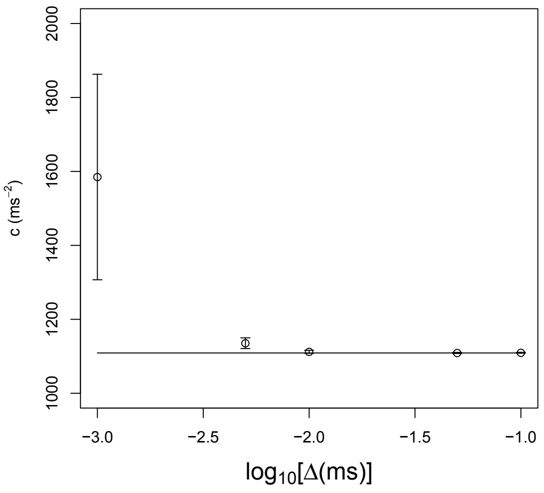


FIG. 5. *Effect of sampling resolution Δ on estimation of c_1 for the single-mode GLE example in Figures* 3 *and* 4. *The horizontal line represents the true value of $c_1 = 1.109 \times 10^3 ms^{-2}$, while the error bars represent* 95% *confidence intervals, which are symmetric about the estimates represented by a hollow point.*

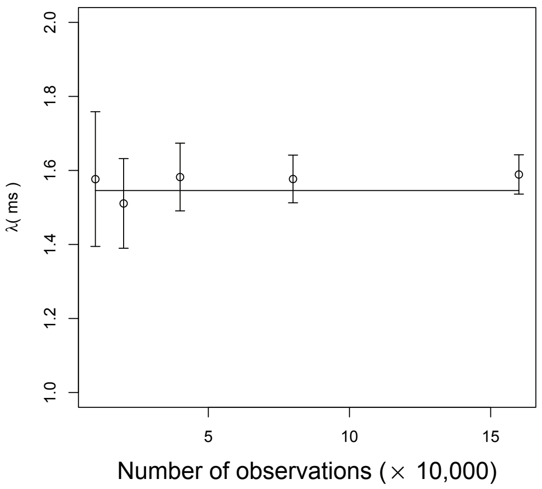


FIG. 6. *Parameter estimation as a function of the number of observations for the single-mode GLE of Figures* 3–5. *The sampling interval is fixed, $\Delta = 0.1ms$, which is a good sampling rate to estimate $\lambda_1 = 1.5ms$, as shown in Figure* 4. *The horizontal line represents the true value of λ_1, and the error bars represent* 95% *confidence intervals which are symmetric about the estimates represented by a hollow point.*

for a given sampling time. This is a general feature of maximum likelihood estimators, and its theoretical verification is given in Appendix B as a consequence of the asymptotic normality of the estimators.

With a single-mode exponential kernel, the quadrature solution of the GLE can be extended to an explicit formula for ensemble averages, particularly for autocorrelations of velocity and displacement (cf. [7]). We drop the subscript 1 on all parameters for these 1-mode formulas. The velocity autocorrelation is given by

$$\langle V(t)V(t')\rangle = \frac{k_BT}{m\beta(1-\beta)}e^{-\frac{1-\beta}{2\lambda}|t-t'|} - \frac{k_BT}{m\beta(1+\beta)}e^{-\frac{1+\beta}{2\lambda}|t-t'|}, \tag{3.17}$$

while the MSD is

$$\begin{aligned}\langle[X(t)-X(t')]^2\rangle = \frac{4k_BT}{m}\Bigg\{&\frac{2\lambda}{1-\beta^2}|t-t'| - \frac{2\lambda^2(3+\beta^2)}{(1-\beta^2)^2}\\ &+ \frac{\lambda^2}{\beta(1-\beta^2)^2}\left(e^{-\frac{1-\beta}{2\lambda}|t-t'|}(1+\beta)^3 - e^{-\frac{1+\beta}{2\lambda}|t-t'|}(1-\beta)^3\right)\Bigg\},\end{aligned} \tag{3.18}$$

where $\beta = \sqrt{1-4c\lambda^2}$ and $c = 6\pi aG/m$ from (3.5). For sufficiently short times, the MSD (3.18) exhibits ballistic behavior, $\langle[x(t)-x(0)]^2\rangle \approx k_BTt^2/m$, and for sufficiently long times, diffusive scaling emerges, $\langle[x(t)-x(0)]^2\rangle \approx 2k_BTt/m\lambda c$. For intermediate times, a power law fit of the MSD yields a range of exponents depending on the window in which one chooses to fit.

We note that the parameter β can be purely imaginary, as pointed out in [7], which is clear from the formula (3.18). Oscillations are predicted in the velocity correlation and MSD whenever physical parameters obey $4c\lambda^2 > 1$. When extended to the more general case of multiple exponentials, similar oscillations appear since the relevant matrix A often has a pair of complex eigenvalues.

This GLE model phenomenon predicts high frequency (short time) oscillations in experimental path data, *even after ensemble averaging of path time series*, which translates to a source of high frequency error of MSD in experimental measurements because of the phase mismatch between these inherent oscillations and experimental sampling time. We do not know if this property is generic for a wider class of kernels.

3.3. GLE model illustration with a 4-mode Rouse kernel.

A classical model due to Rouse (cf. [5]) yields a special class of M-mode kernels for which GLE diffusive transport properties are explicitly solvable. A 4-mode Rouse kernel is implemented now to further illustrate the direct and inverse tools, and to benchmark our direct simulations against exact MSD scaling laws. To construct a Rouse kernel, polymer chains are divided into spherical mass segments connected by linear springs of equilibrium length b (beads in polymer chain); and a kernel function of a series of exponentials with same weight and different characteristic time is then followed [4, 17]. A Zimm kernel, in which a different exponential spectra is derived, is presented next. More complex molecular models may incorporate overlap and entanglements of polymer chains, or even chemical interactions between Brownian particles and local environment. Our focus in this paper is to model the fluctuations without attempting to dissect the various sources. Our goals in this example are once again as follows: for inversion, to find the best GLE kernel to fit measured path data; for direct prediction, to simulate particle paths or the statistics of paths for a known prescribed GLE kernel.

To prescribe the kernel for a Rouse chain solution, each segment in a polymer chain is assigned friction coefficient ξ_b, and the weight and characteristic times for the exponentials of the ith mode are given by (with N_m the number of segments in a

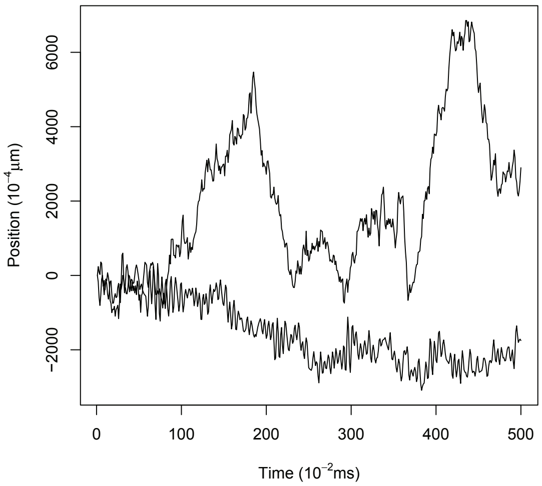


FIG. 7. *Sample discrete AR simulation for a GLE with a* 4*-mode Rouse kernel (top path) compared to a Brownian motion (Langevin equation path) with the same local variance.*

polymer chain)

$$G_i = G_0 = \nu k_B T, \quad \lambda_i = \frac{\xi_b}{16 k_B T \beta_b^2 \sin^2(i\pi/2(N_m+1))}, \tag{3.19}$$

where ν is the number density of polymer chains and $\beta_b = 3/(N_m b^2)$. In the example to follow, we choose $\nu = 2\%$. We now specify all parameter values in the 4-mode Rouse-GLE model. The passive tracer bead is 1μm in diameter of mass $m = 1.05 \times 10^{-9}$mg. The single weight factor is given by $G_0 = G = 1.035 \times 10^{-5}mg/ms^2\mu$m, so that our rescaled parameters are $c = c_i = 6\pi a G_0/m = 4.440 \times 10^{-4}(\text{ms})^{-2}$. The Rouse relaxation times are, from (3.19), $\lambda_1 = .02415$, $\lambda_2 = .04294$, $\lambda_3 = 0.09661$, and $\lambda_4 = .38643$ in units of ms. Figure 7 shows a typical time series for particle position for this GLE-Rouse kernel, extracted from the full vector AR simulation. For comparison, we have included a sample path for a random walk with independent steps. The variance of the steps for both time series are the same; therefore, the figure gives a clear illustration of the effect of dependency alone in suppressing the diffusion of a particle.

We simulate 200 paths with sampling time $\Delta = 10^{-3}$ms for 10^4 steps. Figure 8 shows the ACF (MSD) for the position of the paths, computed by ensemble averaging of the 200 paths (green dots). This result is compared with the analytical scaling law (yellow dashed curve) for a Rouse chain [4, 17]. (Later in this section, we present a more general result from vector Langevin stochastic processes: an explicit quadrature formula for the autocorrelation matrix of the vector Langevin process. This formula allows one to bypass single paths and ensemble averaging of them to directly simulate MSD and velocity autocorrelations.) Note that the MSD starts out with ballistic scaling for times far below the shortest relaxation time and eventually becomes diffusive for times longer than the largest relaxation time. Subdiffusive scaling occurs between the shortest ($t = 0.02415$ms) and longest ($t = 0.38643$ms) relaxation times, consistent with Rouse behavior.

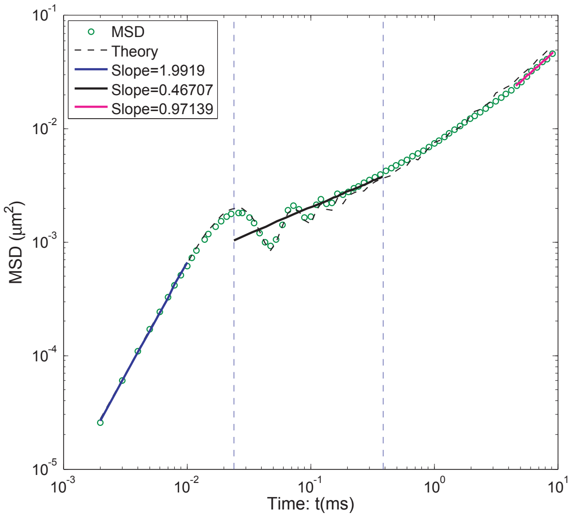


FIG. 8. *MSD of GLE sample paths for a 4-mode Rouse diffusive transport modulus.* 200 *paths are generated for a* 1 μm *diameter bead at* $293K$*. The Rouse relaxation times are* $\lambda_1 = .02415$*,* $\lambda_2 = .04294$*,* $\lambda_3 = 0.09661$*, and* $\lambda_4 = .38643$ *in units of* ms*, with equal weights for each mode,* $G_0 = 1.035 \times 10^{-5} mg/ms^2\mu m$*. To benchmark analytical scaling laws, a linear fit between the two vertical dashed lines (from the shortest to longest relaxation times) confirms the MSD power law of* 0.5 *for the Rouse model. The short term ballistic and long term diffusive scaling are also confirmed.*

Now we turn to the application of inverse methods for the path data, treating the data as though it were generated experimentally. To reveal the effective memory in this system, we first "preprocess" one sample time series to get an estimate of the ACF for velocity, which is obtained by differencing the position data. We use this proxy for the ACF of velocity to obtain initial conditions for the maximum likelihood method of fitting memory kernels. The ACF result is shown in Figure 9. Note the oscillatory behavior of the ACF, clearly indicating that the process is not consistent with a particle diffusing in a purely viscous fluid. (This remark relates to the earlier analysis of oscillations that arise in single-mode GLE models, which persist for this Rouse kernel.)

The ACF in this context is being used as an exploratory tool to gauge the amount of dependency present in the data before using the maximum likelihood techniques to fit the model. The ACF gives a proxy here for the longest relaxation time seen in the data which gives an initial guess for the single-mode model. If no significant lags were seen, then it is likely that all relaxation times are below the sampling rate and more frequent observations are necessary to estimate relaxation times. If the researcher suspects well-separated relaxation times over several orders of magnitude, then one could use more coarsely sampled data to fit the longest times and after fitting use a finer grid to fit shorter relaxation times. The ACF can be used to guide these explorations of widely separated times.

In general, the number of exponential modes that best fit the underlying process that generated the data is not known. The strategy begins by positing a single exponential to fit the data, from which the ACF produces a rough guess of 0.04 ms for the relaxation time. Our experience with numerical and experimental data indicates that fitting the data to a single-mode kernel tends to be quite stable, and this initial step consistently gives the same results independent of the initial guess for the

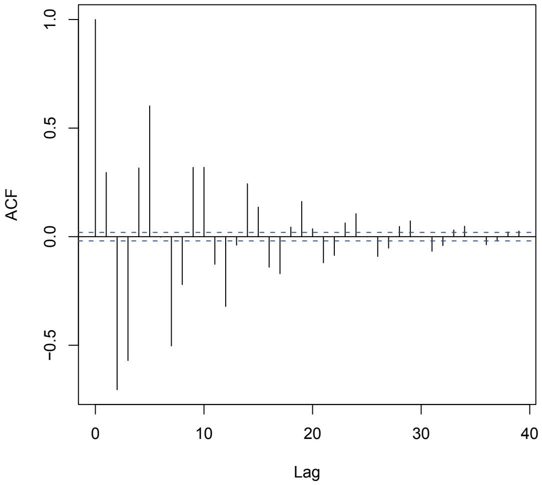


FIG. 9. *ACF for velocity approximated by differencing of position data for the discrete AR process corresponding to a GLE with the* 4*-mode Rouse kernel of Figure* 8*.*

relaxation time. The estimated parameter values are $\hat{\lambda}_1 = 5.519 \pm 0.071(10^{-2}\text{ms})$ and $\hat{c} = 1.77 \pm 0.003(10^{-3}\text{ms}^{-2})$. Not surprisingly the estimated value of c is almost exactly four times the true value since the data was generated from a 4-mode model. (Fitting a single-mode model is essentially the same as fitting a 4-mode model where all the modes have the same relaxation time, thus yielding a $\hat{c}$ that is roughly four times the true value.)

We would like to be able to assess the quality of the fits being performed. One diagnostic tool for investigating how well the model predicts the data is *the ACF of the residuals*. This is shown in Figure 10. If the model has successfully captured all the dependencies in the data, then we expect the ACF of the residuals to be consistent with white noise. Note that the first few lags show a significant negative correlation, indicating that the single-mode model cannot account for all the dependency in the data.

We proceed to a 2-mode kernel which requires initial guesses for each relaxation time. If $\hat{\lambda}_1$ is the estimate for the single-mode case, one reasonable approach is to use $\hat{\lambda}_1 \pm \hat{\lambda}_1/2$ as the initial guesses for the 2-modes. In this way, each time we add an additional mode to the model, we split the longest relaxation time and use the estimates obtained from fitting the previous model as an initial guess for the remaining relaxation spectra. That is, for an M-mode model, our initial guesses for the λ's will consist of the $(\hat{\lambda}_1, \ldots, \hat{\lambda}_{M-2})$ obtained by fitting an $M-1$ model, and for the two longest relaxation times we use $\lambda_{M-1} = \hat{\lambda}_{M-1} - (\hat{\lambda}_{M-1} - \hat{\lambda}_{M-2})/2$ and $\lambda_M = \hat{\lambda}_{M-1} + (\hat{\lambda}_{M-1} - \hat{\lambda}_{M-2})/2$. Therefore, for the 2-mode model, we choose initial conditions of 0.0275ms and 0.0825ms for the λ's and use $\hat{c}$ from the single-mode model as the initial condition for c. This produces $\hat{\lambda}_1 = 3.023 \pm 0.043(10^{-2}\text{ms})$, $\hat{\lambda}_2 = 19.30 \pm 0.73(10^{-2}\text{ms})$, and $\hat{c} = 0.886 \pm 0.001(10^{-3}\text{ms}^{-2})$. In this case the estimate for c is roughly twice the true value.

The ACF for the residuals of the 2-mode fit (not shown) indicates that we have captured most of the dependencies in the data. Figure 11 shows a plot of the sum of

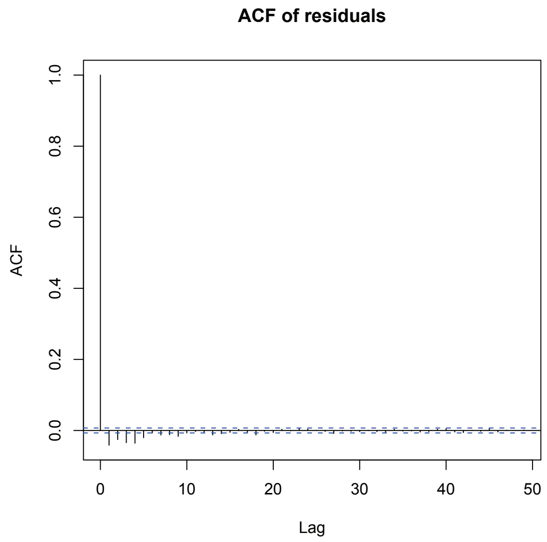


FIG. 10. *ACF of residuals for fitting a single-mode GLE kernel to data generated from a discrete AR process with a* 4*-mode kernel.*

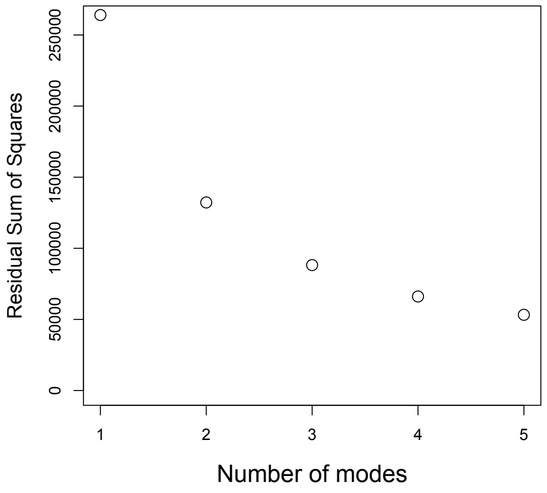


FIG. 11. *The sum of squared residuals when fitting kernels with* 1–5 *modes to* 4*-mode data.*

the squared residuals as a function of the number of modes used to fit the data. Note that there is a large reduction in the sum of the squared residuals in going from one to two modes, but there is no evidence of convergence yet.

We next fit a 3-mode kernel. Using the method described above, the initial guesses for the λ's (in 10^{-2}ms) are 3.023, 11.0, and 27.0. The estimated values for the relaxation times (in 10^{-2}ms) are $\hat{\lambda}_1 = 2.525 \pm 0.060$, $\hat{\lambda}_2 = 7.020 \pm 0.461$, and $\hat{\lambda}_3 = 25.50 \pm 1.99$, and the estimate of c is $\hat{c} = 0.592 \pm 0.001(10^{-3}\text{ms}^{-2})$.

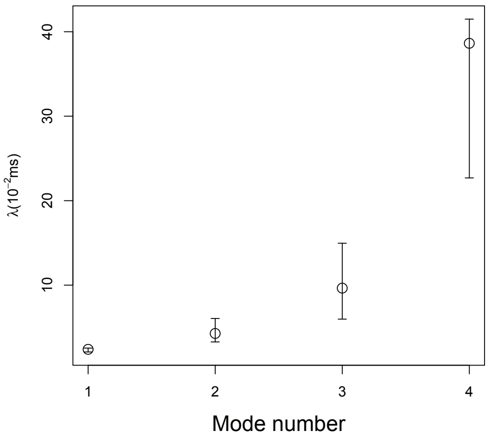


Fig. 12. *Proof-of-principle: Maximum likelihood recovery of a* 4*-mode Rouse relaxation spectrum from numerical time series data. The error bars are symmetric about the estimate, with the open circles being the true values.*

As expected, the estimated value of c is roughly $4/3$ the true value. Note that there is still a significant drop in the sum of the squared residuals (Figure 11). Figure 12 shows results for the estimated values of the relaxation times when a 4-mode kernel is used. For this case the initial guesses for the λ's (in 10^{-2}ms) are 2.525, 7.02, 16.0, and 43.0. Notice that the true λ values all lie within the error bars. For c, we obtain an estimate of $0.443622 \pm 0.00074(10^{-3}\text{ms}^{-2})$, which is very close to the true value.

Attempting to fit a 5-mode kernel with initial guesses of $\lambda_i = 2.322$, 4.670, 10.47, 21.0, and 43.0 (in units of (10^{-2}ms), we obtain estimates for the λ's of 2.179, 3.748, 7.23, 14.947, and 33.897 (in 10^{-2}ms). However, the estimated covariance matrix has negative values on the diagonal, indicating a problem with the maximization process. There is also not a very large reduction in the sum of the squared residuals (Figure 11), which means that the additional parameter does not meaningfully contribute to explaining the data.

While additional parameters will almost always lead to a decrease in the residual sum of squares, it is clear in this case that the fit is unreliable since the approximated covariance matrix is not positive definite. Therefore we conclude that four modes provide an accurate representation of the data.

Next, we perform simulations to gauge the convergence of the parameter estimates with increased data and to test the dependency of the fit on changes in the sampling interval. Figure 13 shows the estimated values of λ_3 and λ_4 as a function of the number of data points in the time series. (The fits for the other two relaxation times are significantly better and omitted for clarity.) The convergence rate appears to be on the order of $n^{-1/2}$, consistent with the earlier derivation of the confidence interval. Figure 14 shows the estimated values of λ_3 and λ_4 as functions of the sampling time Δ. The results are similar to those for the Langevin equation (Figure 2). That is, the method has difficulties estimating the relaxation times if too short or too long a sampling time is used.

3.4. Direct GLE simulations of MSD and velocity autocorrelations. Ensemble average information for vector Langevin equations can be expressed in quadra-

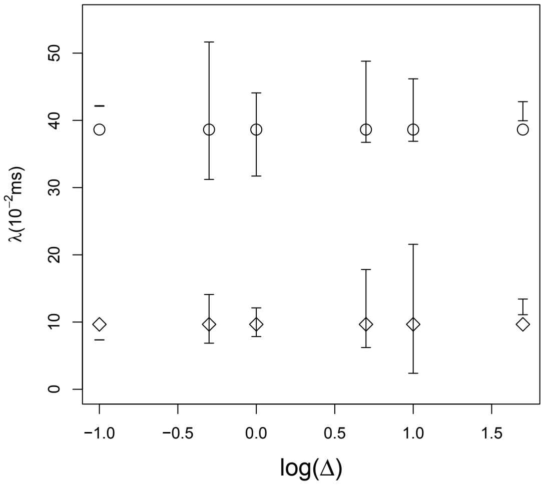


FIG. 13. *Parameter estimation versus sampling rate for the longest relaxation times* λ_3 *and* λ_4 *in a* 4*-mode kernel. The error bars are symmetric about the estimate with the open circles being the true values. The x-axis represents the log of* Δ *(sampling time).*

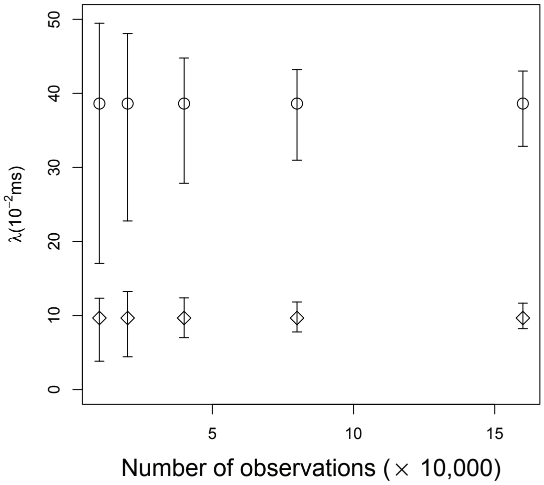


FIG. 14. *Parameter estimation versus number of observations (in units of* 10^4*) for the two longest relaxation times* λ_3 *and* λ_4 *in a* 4*-mode kernel. The error bars are symmetric about the estimate with the open circles being the true values.*

ture form by the appropriate averaging of the exact quadrature formula for individual paths. The full matrix of autocorrelations for a vector Ornstein–Uhlenbeck process is

$$\langle Y(t)Y^T(t')\rangle = \int_0^t \int_0^{t'} ds_1 ds_2 \delta(s_1 - s_2) e^{A(t-s_1)} K K^T e^{A^T(t'-s_2)}. \tag{3.20}$$

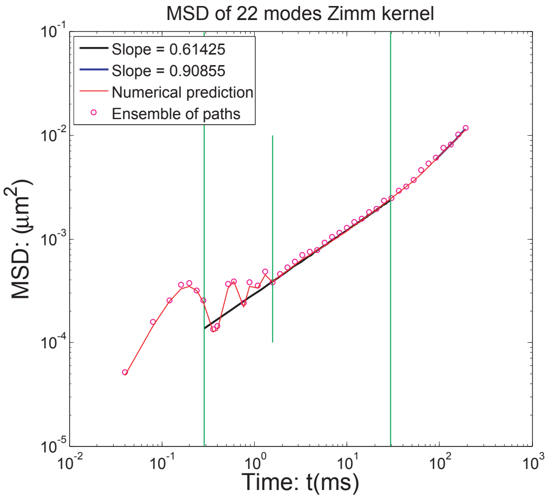


FIG. 15. *MSD of the GLE with a* 22*-mode Zimm kernel. The smallest relaxation time is* 0.2885*ms, the longest is* 29.77*ms; the two vertical lines mark the time span between them, over which a power law of* 0.62 *fairly well approximates the theoretical Zimm model value of* $\frac{2}{3}$.

The (1,1) entry of the resulting matrix gives the MSD, and the (2,2) entry gives the velocity autocorrelation. *The practical ramification of this formula is that one can directly generate statistical properties for a known GLE M-mode diffusive transport modulus without the need to generate sample paths and then take ensemble averages.* For the special case of a single-mode exponential kernel, the integral representation can be solved explicitly, which gives the result presented earlier in (3.17), (3.18).

In Figure 8 for the 4-mode Rouse kernel, the MSD is computed two ways: from averaging of 200 sample paths generated from the GLE model and depicted by (blue) circles; and then directly from the autocovariance formula (3.20) and depicted by the (yellow) dashed line. Figure 8 convincingly reproduces the correct MSD power law behavior of Rouse theory, namely an exponent of $\frac{1}{2}$ when fitted over intermediate times between the relaxation spectra. This comparison provides another benchmark on the direct simulation tools, both for sample paths and for the autocovariance of GLE processes.

We now illustrate that the methods are not "mode limited," by running direct simulations for beads of the same size and mass as in Figure 8, but with a GLE diffusive transport modulus specified by a 22-mode Zimm kernel. The model posits 1100 monomers along each polymer chain, which we divide into 22 subunits, which gives 22 modes and an explicit relaxation spectrum. Figure 15 shows the MSD statistics, again generated both by ensemble averaging of paths and by the ACF (3.20). The simulations predict an MSD power law scaling exponent of 0.62 when fitted between the shortest and longest relaxation spectra, which reasonably approximates the $\frac{2}{3}$ theoretical value of the Zimm model.

3.5. Comparison with the MW inverse method. The inverse characterization framework for the memory kernel proposed in this paper focuses on single path information in the time domain, which is a complement to the transform space formulation of Mason and Weitz [10, 11, 12]. We now compare the two approaches on data generated by the GLE with the 4-mode Rouse kernel above. To make a fair

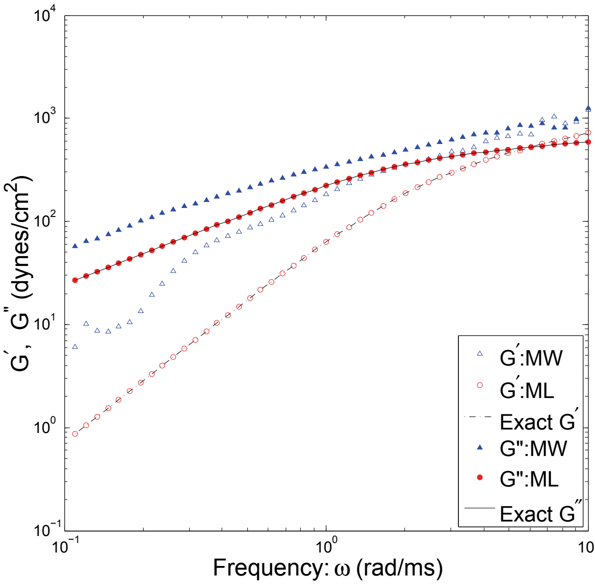


Fig. 16. *The real ($G'(\omega)$) and imaginary ($G''(\omega)$) parts of the transform of the GLE memory kernel, recovered from the same numerical GLE data with a* 4*-mode Rouse kernel, by the maximum likelihood and MW methods. The maximum likelihood results correspond to a best* 4*-mode exponential kernel fit.*

comparison, we simulate an experiment which gathers many bead paths.

In Mason and Weitz's original contribution [12], the memory kernel is transformed to frequency space following the standard definitions and notations of linear viscoelasticity [5]. The unilateral Fourier transform of the kernel, recalled below, is assumed to be proportional to the frequency-dependent shear viscosity function: $m\xi^* = 6\pi a\eta^*$, which is the generalized Stokes–Einstein relation. Recall that the transformed shear modulus G^* is proportional to the transformed viscosity function, $G^* = i\omega\eta^*$, i.e.,

$$G^*(\omega) \equiv i\omega \int_0^\infty \eta(t) e^{-i\omega t} dt = G^{'} + iG^{''}. \tag{3.21}$$

If we now assume the 4-mode Rouse kernel, the corresponding real and imaginary parts of G^* are

$$G'(\omega) = \sum_{i=1}^{4} \frac{G_0 \omega^2 \lambda_i^2}{1+\omega^2\lambda_i^2}, \quad G''(\omega) = \sum_{i=1}^{4} \frac{G_0 \omega \lambda_i}{1+\omega^2\lambda_i^2}, \tag{3.22}$$

where G_0 and λ_i are defined in (3.19).

The "experimental data" consists of 200 paths of 1μm diameter tracer beads, generated from the GLE algorithm described earlier. First we implement the MW method. We calculate the MSD from these 200 paths, shown in Figure 8. Next, the MSD versus t is transformed to the frequency domain, together with the GSER, to arrive at G^* (see [10] for details). We note that the MW method is applied only over the monotone part of the MSD curve in Figure 8, which optimizes the accuracy of the MW reconstruction of $G^*(\omega)$. The results are graphed in Figure 16. Second, we apply the maximum likelihood method to gain the best 4-mode fit to the path data. G^* is then given by (3.22) with the maximum likelihood estimators, graphed in Figure 16. The MW method overestimates G' and G'' in this frequency range.

If we further wanted to invert $G^*(\omega)$ to recover $G(t)$, clearly the maximum likelihood method requires no work. From the MW estimate $G^*(\omega)$, we refer to [14, 20] for numerical strategies to estimate $G(t)$, including an exponential fit.

We comment that this comparison is made on data for which our methods are designed to do well. The real test, on experimental data, remains for future comparisons.

4. Conclusions. A time-domain statistical strategy has been developed for passive microbead rheology which serves two purposes: as an inversion toolkit for recovery of the diffusive transport modulus in a generalized Langevin equation from experimental time series, and as a direct simulation toolkit for pathogen diffusion of single particles and statistical correlations if the diffusive transport modulus is known. These direct and inverse algorithms combine to a general package for anomalous diffusive transport of pathogens in soft matter, which we anticipate to be complementary to the MW experimental and theoretical protocol [10, 11, 12]. These tools are presently being applied to characterization of pulmonary liquids with our colleagues Superfine, Hill, and Cribb in the Virtual Lung Project at UNC.

We mention another related approach based on fractional Brownian diffusion developed by Kou, Xie, and coworkers [9, 13]. The approach taken in that work is to formulate the GLE using fractional Brownian white noise as the stochastic driving force. A benefit of this formulation is that the number of parameters is limited; the modeling feature that is distinct from our methods is that the autocovariance function decays as a specific power law uniformly in time. If MSD experimental data reflects a uniform power law scaling over the experimental time series, then the fractional Brownian diffusion model should be strongly considered. The method of fitting relies on estimating the autocovariance function for velocity and then fitting the parameterized autocovariance to this estimated function. Standard errors may then be obtained via simulation. The drawbacks include stochastic approximation in the simulation methods and the difficulty in estimating the autocovariance of the velocity when only position is observed. Our method overcomes these difficulties but is limited to models consistent with autocovariance functions which for long lags have an exponential decay. Our formulation also allows for a greatly simplified simulation method and a maximum likelihood parameter estimation procedure which may use experimental data more efficiently.

An open question relates to the range and timescales of power law behavior in the MSD that are possible for GLE models with the class of M-mode exponential kernels considered in this paper. So far, we have reproduced the classical Rouse and Zimm MSD scalings on intermediate timescales between the shortest and longest relaxation times for kernels with the Rouse and Zimm relaxation spectra. Our preliminary numerical studies show that a wide range of power law behavior is possible as the relaxation spectrum and the respective weights for each mode are varied, and recent, unpublished analytical results of Scott McKinley at Duke University confirm this numerical evidence.

These tools are viewed as a foundation for further extensions of the single-bead and two-bead models and experiments. The ability to separate local bead-fluid interactions from the bulk viscoelastic modulus, and to identify heterogeneity from single-particle and two-particle statistical correlations, are key future applications of these tools.

Appendix A. The Kalman filter. Similar discussions to the following, based on [19], may be found in numerous texts [8, 2]. The framework of the Kalman filter is to take a linear system model and an observation model which depends linearly on the state of the system. We call this general setup a linear state space model and use the following notation: The system equation is

$$Y_n = BY_{n-1} + \epsilon_n, \tag{A.1}$$

where $\epsilon_n \sim N(0, S)$, and the observation equation is

$$U_n = CY_n + \xi_n, \tag{A.2}$$

where $\xi_n \sim N(0, D)$. Also, note that ϵ_n and ξ_n are independent sequences and independent of each other. (Here we have included an error term for U_n which is the case in the standard Kalman filter. In the present paper, we assume no observation error, and so the D matrix will be zero.)

The goal of the Kalman filter is to calculate the conditional distribution of Y_n, given the observations $U_1, \ldots, U_n$. The mean of this conditional distribution is an estimate (which is optimal in certain ways) of Y_n. We are estimating the "hidden" elements of the process by conditioning on the observed elements of this process. For this procedure to be computationally feasible, a recursive algorithm is necessary. In other words, we would like to calculate the new conditional distribution of Y_n given $U_1, \ldots, U_n$ using only the conditional distribution of Y_{n-1} given $U_1, \ldots, U_{n-1}$ and a new observation U_n.

As a preliminary, the calculations of the Kalman filter rely on a basic theorem from multivariate statistical analysis which allows us to calculate the distribution of a portion of a Gaussian random vector conditioned on the other portion. For a normal random vector, A,

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right], \tag{A.3}$$

we have that the distribution of A_1, given that $A_2 = a$, is

$$\mathcal{N}[\mu_1 + \sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]. \tag{A.4}$$

This also works in reverse—if $A_2 \sim \mathcal{N}[\mu_2, \Sigma_{22}]$ and the distribution of A_1 is given in (A.4), then the joint distribution is given by (A.3). (We are using the notation $\mathcal{N}[\mu, \Sigma]$ for multivariate normal distribution with mean vector μ and covariance matrix Σ.)

As mentioned, we would like to find a set of recursive equations such that if we had the new observation U_n and the distribution of $Y_{n-1}|U_1, \ldots, U_{n-1}$ (which we write as $Y_{n-1|n-1}$ throughout), then we can find the distribution $Y_{n|n}$. This distribution is the Kalman filter at time n. So, let's assume that we have the conditional distribution of $Y_{n-1|n-1}$, where we call the conditional mean of this random vector $\hat{Y}_{n-1}$ and the conditional covariance P_{n-1}. Now, using (A.1), we can calculate the distribution for $Y_{n|n-1}$ which will be

$$Y_{n|n-1} = \mathcal{N}[B\hat{Y}_{n-1}, BP_{n-1}B^t + S]. \tag{A.5}$$

For simplicity, we use the notation $R_{n-1} = BP_{n-1}B^t + S$ for the covariance matrix. Combining (A.2) and (A.5) yields

$$\begin{pmatrix} U_{n|n-1} \\ Y_{n|n-1} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} CB\hat{Y}_{n-1} \\ B\hat{Y}_{n-1} \end{pmatrix}, \begin{pmatrix} D + CR_{n-1}C^t & CR_{n-1} \\ R_{n-1}C^t & R_{n-1} \end{pmatrix}\right]. \tag{A.6}$$

Right now, we need only to condition $Y_{n|n-1} = Y_n|(U_1, \dots, U_{n-1})$ on $U_{n|n-1} = U_n|(U_1, \dots, U_{n-1})$ to give us $Y_{n|n} = Y_n|(U_1, \dots, U_n)$, which is what we want. Another application of the theorem gives us that the mean of $Y_{n|n}$ is

$$\hat{Y}_n = B\hat{Y}_n + R_{n-1}C^t(D + CR_{n-1}C^t)^{-1}(U_n - CB\hat{Y}_{n-1}) \tag{A.7}$$

and the covariance is

$$P_n = R_{n-1} - R_{n-1}C^t(D + CR_{n-1}C^t)^{-1}CR_{n-1}. \tag{A.8}$$

So, we have derived the necessary recursions to take a new observation at time n and the filter at time $n-1$ (i.e., the distribution of Y_{n-1}, given the observations up to time $n-1$) and obtain the value of the filter at time n.

For our application, one element is particularly important—the one-step prediction for the observation process which is the distribution of U_n given $U_1, \dots, U_{n-1}$, i.e., $U_{n|n-1}$. This is given, however, in the first entry of the combined vector on the left-hand side of (A.6). Explicitly,

$$U_{n|n-1} \sim \mathcal{N}[CB\hat{Y}_{n-1}, D + CR_{n-1}C^t]. \tag{A.9}$$

This calculation is used in the error-prediction decomposition approach to calculating the likelihood function.

Appendix B. Asymptotic normality of maximum likelihood estimators. A key benefit of the maximum likelihood method is the ability to calculate standard errors on the estimates. In general, one starts with a model that depends on the parameters Θ, and then maximizes the likelihood function with respect to the model parameters to obtain the best estimate $\hat{\Theta}$ for the parameters. Under certain conditions, $\sqrt{N}(\hat{\Theta} - \Theta)$ converges to a multivariate normal with mean zero and covariance matrix $I^{-1}(\Theta)$, where $I(\Theta)$ is the information matrix [8] given as

$$I(\Theta) = -E\nabla^2 \log L(\Theta). \tag{B.1}$$

The necessary conditions that need to be satisfied are the following:

1. $I^{-1}(\Theta)$ must be positive definite.
2. $\hat{\Theta}$ must be in the interior of the parameter space.
3. $\log L(\Theta)$ has third-order continuous derivatives in the neighborhood of the true parameter values Θ.
4. Θ is identifiable. In other words, for each set of data, $L(\Theta)$ is a one-to-one function of Θ.

We approximate $I^{-1}(\Theta)$ by finding the Hessian of the logarithm likelihood function numerically with respect to the parameters evaluated at the maximum.

Appendix C. Evaluation of autocovariance. We discuss how the covariance matrix S for a GLE with M-mode kernel in (2.17), while $(2M+2) \times (2M+2)$ coefficient matrices A and K are defined as in (3.15), can be numerically calculated accurately and efficiently. The only difficulty is in finding all $2M+2$ eigenvalues of A; the remaining steps are straightforward.

C.1. Calculation of eigenvalues. For simplicity, we introduce parameters

$$c_i = \frac{6\pi a G_i}{m} = \frac{6\pi a \eta_i}{m\lambda_i}, \quad \sigma_i = \sqrt{\frac{k_B T}{m}}, \quad \kappa_i = \sqrt{\frac{2c_i}{\lambda_i}}. \tag{C.1}$$

Clearly, M eigenvalues, $\{-1/\lambda_i\}_{i=1}^{M}$, are easy to get. The remaining $2M+2$ are determined by the roots of the polynomial equation

$$P_d(x) = x\left(x\prod_{i=1}^{M}\left(x+\frac{1}{\lambda_i}\right)+\sum_{i=1}^{M}c_i\prod_{j\neq i}^{M}\left(x+\frac{1}{\lambda_j}\right)\right)=0. \tag{C.2}$$

First we factor out the simple zero eigenvalue associated with the position equation and then consider the remaining $M+1$ eigenvalues by studying the roots of the polynomial equation

$$P(x) = x\prod_{i=1}^{M}\left(x+\frac{1}{\lambda_i}\right)+\sum_{i=1}^{M}c_i\prod_{j\neq i}^{M}\left(x+\frac{1}{\lambda_j}\right)=0. \tag{C.3}$$

If we rewrite the above polynomial (C.3) by dividing it with $\prod_{i=1}^{M}(x+1/\lambda_i)$, we have a new function

$$Q(x) = x + \sum_{i=1}^{M}\frac{c_i}{x+1/\lambda_i}, \tag{C.4}$$

which has the same roots as $P(x)$. Recall $0<\lambda_1<\cdots<\lambda_M$. Clearly $Q(x)$ changes sign, and therefore has one zero, in each interval $(-1/\lambda_i,-1/\lambda_{i+1})$. These are easily found by iteration. This yields $M-1$ eigenvalues, denoted $\{x_i\}_{i=1}^{M-1}$, and only two remain.

The polynomial $P(x)$ of (C.3) has the form

$$P(x) = (x^2+bx+d)\prod_{i=1}^{M-1}(x-x_i)=0, \tag{C.5}$$

where d and b are given explicitly from $\{-1/\lambda_i\}_{i=1}^{M}, \{x_i\}_{i=1}^{M-1}$:

$$\begin{aligned} d &= \frac{P(0)}{\prod_{i=1}^{M-1}(-x_i)} = \frac{\sum_{i=1}^{M}c_i\prod_{j\neq i}^{M}\frac{1}{\lambda_j}}{\prod_{i=1}^{M-1}(-x_i)} = \frac{\sum_{i=1}^{M}c_i\prod_{j\neq i}^{M}\frac{1}{\lambda_j}}{\prod_{i=1}^{M-1}|x_i|} > 0,\\ b &= \frac{\prod_{i=1}^{M}(1+\frac{1}{\lambda_i})}{\prod_{i=1}^{M-1}(1-x_i)} + \frac{\sum_{i=1}^{M}c_i\prod_{j\neq i}^{M}(1+\frac{1}{\lambda_j})}{\prod_{i=1}^{M-1}(1-x_i)} - 1 - d > 0. \end{aligned} \tag{C.6}$$

This completes the calculation of all $2M+1$ eigenvalues, and we note that the last two roots have negative real part due to $b>0$. If the last two roots are complex conjugates, then the matrix A is diagonalizable only in the complex space.

Similarly, for the matrix As in (2.17), where s is a scalar, all the eigenvalues scale explicitly with s and the eigenvectors remain the same.

For $M=1,2,3$, there are analytical formulas for the roots of the polynomial. In the single-mode case, $M=1$, the eigenvalues are

$$\omega_1 = -\frac{1}{\lambda},\quad \omega_2 = -\frac{1}{2}\left(\frac{1}{\lambda}+\sqrt{\frac{1}{\lambda^2}-4c_1}\right),\quad \omega_3 = -\frac{1}{2}\left(\frac{1}{\lambda}-\sqrt{\frac{1}{\lambda^2}-4c_1}\right),\quad \omega_4 = 0, \tag{C.7}$$

with easily calculated eigenvectors. The covariance matrix S of (2.17) can thus be calculated in closed form.

For general M, from (C.5) and (C.6), fast and efficient numerical schemes could be found for the calculation of eigenvalues and eigenvectors.

C.2. Calculation of the covariance matrix S. Given this detailed spectral information for A, we can precompute the covariance matrix, as shown below.

First we assume that the matrix A has full span of eigenvectors R (its inverse is R^{-1}),

$$A = R\Lambda R^{-1}, \quad A^2 = AA = R\Lambda R^{-1}R\Lambda R^{-1} = R\Lambda^2 R^{-1}, \tag{C.8}$$

where Λ is a diagonal matrix whose diagonal components are the eigenvalues of A.

By definition,

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} = \sum_{n=0}^{\infty} \frac{R\Lambda^n R^{-1}}{n!} = R\left(\sum_{n=0}^{\infty} \frac{\Lambda^n}{n!}\right) R^{-1} = Re^{\Lambda}R^{-1}, \tag{C.9}$$

where $e^{\Lambda} = e^{\Lambda^T}$ is diagonal and the covariance matrix S can be written as

$$\begin{aligned} &S = R\left(\int_0^{\Delta} e^{\Lambda(\triangle - s)} R^{-1}KK^T(R^{-1})^T e^{\Lambda^T(\triangle - s)} ds\right) R^T \\ \overset{\triangle - s = u}{\Longleftrightarrow}\ &S = R\left(\int_0^{\Delta} e^{\Lambda u} C e^{\Lambda u} du\right) R^T, \end{aligned} \tag{C.10}$$

where we define $C = (R^{-1}K)(R^{-1}K)^T$.

Next, we take advantage of the above properties of the matrix A, as follows. Denoting by $e^{\omega_i u}$ the ith diagonal component of the matrix $e^{\Lambda u}$, where w_i is the ith eigenvalue of the matrix A and C_{ij} is the ith row and jth column component of the matrix C, we see (here $(\bullet_{ij})_{M\times M}$ denotes an M-by-M matrix with ith row and jth column component $\bullet_{ij}$)

$$\begin{aligned} e^{\Lambda u} C e^{\Lambda u} &= (C_{ij} e^{w_i u})_{(2M+2)\times(2M+2)} e^{\Lambda u} \\ &= (C_{ij} e^{(w_i + w_j)u})_{(2M+2)\times(2M+2)}. \end{aligned} \tag{C.11}$$

So the covariance matrix admits

$$\begin{aligned} S &= R\left(C_{ij}\int_0^{\Delta} e^{(\omega_i + \omega_j)u} du\right) R^T \\ &= R\left(C_{ij}\frac{e^{(\omega_i + \omega_j)\Delta} - 1}{\omega_i + \omega_j}\right) R^T, \end{aligned} \tag{C.12}$$

and after all the eigenvalues ω_i of A are determined, the integral form of S can be precalculated according to the above result, and the integration of the matrix function can be avoided.

Acknowledgments. The authors acknowledge guidance and productive discussions with several colleagues in the Virtual Lung Project: Ric Boucher, Jeremy Cribb, Bill Davis, David Hill, Christel Hohenegger, Richard McLaughlin, Michael Rubinstein, John Sheehan, and Richard Superfine. The authors also acknowledge valuable conversations with Dave Weitz, Victor Breedveld, Eric Furst, Jingfang Huang, and Scott McKinley.

REFERENCES

[1] B. J. Berne, J. P. Boon, and S. A. Rice, *On the calculation of autocorrelation functions of dynamical variables*, J. Chem. Phys., 45 (1966), pp. 1086–1096.

[2] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer, New York, 1991.

[3] P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*, Cambridge University Press, Cambridge, UK, 1995.

[4] M. Doi and S. F. Edwards, *The Theory of Polymer Physics*, Oxford University Press, London, 1986.

[5] J. D. Ferry, *Viscoelastic Properties of Polymers*, 3rd ed., Wiley, New York, 1994.

[6] G. Gardner, A. C. Harvey, and G. D. A. Phillips, *Algorithm AS* 154: *An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering*, Appl. Statist., 29 (1980), pp. 311–322.

[7] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, Academic Press, New York, 1986.

[8] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK, 1989.

[9] S. C. Kou and X. S. Xie, *Generalized Langevin equation with fractional Gaussian noise: Subdiffusion within a single protein molecule*, Phys. Rev. Lett., 93 (2004), paper 180603.

[10] T. G. Mason, *Estimating the viscoelastic moduli of complex fluids using the generalized Stokes-Einstein equation*, Rheo. Acta, 39 (2000), pp. 371–378.

[11] T. G. Mason, H. Gang, and D. A. Weitz, *Diffusing wave spectroscopy measurements of viscoelasticity of complex fluids*, J. Opt. Soc. Amer. A, 14 (1997), pp. 139–149.

[12] T. G. Mason and D. A. Weitz, *Optical measurements of the linear viscoelastic moduli of complex fluids*, Phys. Rev. Lett., 74 (1995), pp. 1250–1253.

[13] W. Min, G. Luo, B. J. Cherayil, S. C. Kou, and X. Sunney Xie, *Observation of a power-law memory kernel for fluctuations within a single protein molecule*, Phys. Rev. Lett., 94 (2005), paper 198302.

[14] S. M. F. D. Syed Mustapha and T. N. Phillips, *A dynamic nonlinear regression method for the determination of the discrete relaxation spectrum*, J. Phys. D Appl. Phys., 33 (2000), pp. 1219–1229.

[15] B. Øksendal, *Stochastic Differential Equations*, Springer, New York, 1998.

[16] S. A. Rice and P. Gray, *Statistical Mechanics of Simple Liquids*, Wiley, New York, 1965.

[17] M. Rubinstein and R. H. Colby, *Polymer Physics*, Oxford University Press, London, 2003.

[18] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, 2nd ed., Springer, New York, 2006.

[19] R. L. Smith, *Time Series*, course notes, Department of Statistics, University of North Carolina, Chapel Hill, NC, 1999.

[20] M. J. Solmon and Q. Lu, *Rheology and dynamics of particles in viscoelastic media*, Current Opinion in Colloid & Interface Sci., 6 (2001), pp. 430–437.

[21] F. Brochard Wyart and P. G. de Gennes, *Viscosity at small scales in polymer melts*, Euro. Phys. J. E, 1 (2000), pp. 93–97.

[22] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, London, 2001.

[23] R. Zwanzig and M. Bixon, *Hydrodynamic theory of the velocity correlation function*, Phys. Rev. A, 2 (1970), pp. 2005–2012.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1309–1333

DIFFRACTION BY A SEMI-INFINITE INTERFACIAL CRACK SANDWICHED BETWEEN TWO ISOTROPIC HALF PLANES*

V. KUBZINA[†‡], A. K. GAUTESEN[§], AND L. JU. FRADKIN[†]

Abstract. This paper addresses the canonical two dimensional problem of diffraction of the plane wave by a semi-infinite interfacial crack sandwiched between two isotropic solids. We restrict ourselves to a ubiquitous case of solids whose contact boundary does not support the Stoneley wave. Its solution can be used in applications to model diffraction from curved cracks with curvature that is small compared to a wavelength.

Key words. diffraction, interfacial crack, semianalytical approach

AMS subject classifications. 45E10, 78A45, 74J20

DOI. 10.1137/070711517

1. Introduction.

Finding a semianalytical solution to the problem of diffraction by a semi-infinite crack sandwiched between two different solids is a well-known problem in the mathematical theory of diffraction. In this paper we consider a two dimensional case involving two different isotropic half planes. Apart from being mathematically challenging, diffraction problems of this kind are of interest in ultrasonic NDE (nondestructive evaluation), particularly because interfacial cracks are often found in laminated composites. It is well known that ultrasonic inspection of such cracks is a challenging engineering problem, and detection of crack tip diffraction is particularly difficult. As a consequence, the defect size can be underestimated. Nevertheless, the advanced phased array transducers offer an improved performance [1], and models of the underlying diffraction process would allow the NDE inspectors to establish whether, in a given configuration, the amplitude of the edge diffracted echoes could exceed the detection threshold [2].

Over the years purely numerical approaches to this kind of problem based on finite differences, finite elements, or boundary integral techniques proved unreliable, because it is difficult to take into account the singularity condition at the crack tip and thus render a solution unique. It is also difficult to keep adjusting numerical schemes to account for different types of wave interaction [3, 4, 5]. Another well-known line of attack is to reformulate the problem in terms of a system of functional equations and to solve those using a numerical Wiener–Hopf factorization technique (see, e.g., [6]). So far, this approach has also met with numerous numerical difficulties and has produced no entirely satisfactory scheme.

*Received by the editors December 20, 2007; accepted for publication (in revised form) October 30, 2008; published electronically February 19, 2009. A related short report is A. K. Gautesen, V. Kubzina, and L. Ju. Fradkin, *Modeling scatter from 2D interfacial crack*, in Review of Progress in Quantitative Nondestructive Evaluation 24A, D. O. Thompson and D. E. Chimenti, eds., Plenum Press, New York, 2005, pp. 35–42.

http://www.siam.org/journals/siap/69-5/71151.html

[†] Waves and Fields Research Group, Electrical, Computer and Communications Engineering Department, Faculty of Engineering, Science and The Built Environment, London South Bank University, London SE1 0AA, UK (vkubzina@tynemarch.co.uk, fradkil@lsbu.ac.uk). The first author had partial financial support from London South Bank University.

[‡] Current address: Tynemarch Systems Engineering Ltd., Crossways House, 54–60 South Street, Dorking, Surrey, RH4 2HQ, UK.

[§] Department of Mathematics, Iowa State University, and Ames Laboratory, Ames, IA 50011 (gautesen@scl.ameslab.gov).

In this paper we follow a semianalytical approach of [7, 8, 9, 10], developed to derive a solution of the elastodynamic equations formulated in terms of displacements, which it reduces to a system of regular integral equations for their Fourier transforms. We start with the elastodynamic integral equation for the displacement based on Green's formula and the extinction theorem for each isotropic half plane and then use operations of dilatation (or divergence) and rotation (or curl) to separate transverse and longitudinal motions. Since the incident wave can be considered as radiated by a line load, we represent the two dimensional free space Green's tensors in terms of the Hankel function of the first kind of the zeroth order and its derivatives. Using the boundary conditions, the Fourier transform of the elastodynamic integral equation is reduced to a system of four functional equations in eight "half unknowns." Then, the problem is reformulated in terms of traction and crack opening displacement, both of which can be decomposed into singular and nonsingular parts. The singular parts relate to the well-known geometricoelastodynamic (GE) body waves. The nonsingular parts constitute new unknowns. By using a Hilbert-type integral transform, the functional equations are transformed into four regular integral equations in four unknowns. In turn, these are solved numerically. In the far field, diffraction body wave coefficients are obtained. The method can be generalized to model transversely isotropic media.

2. The problem statement.

We consider a two dimensional semi-infinite crack (see Figure 2.1) sandwiched between two different isotropic media $I^{(j)}$, where superscript $j = 1$ corresponds to the medium occupying the "upper" half plane and $j = 2$ means the "lower" half plane. Let the crack be irradiated by a longitudinal ($n = 1$) or transverse ($n = 2$) plane wave, which is incident from the medium $I^{(m)}$, $m = 1$ or 2, and propagates there with the speed $c_n^{(m)}$, where m and n are both fixed throughout the paper. Further, let us assume without loss of generality that the longitudinal speed $c_1^{(1)}$ in the medium $I^{(1)}$ is greater than the longitudinal speed $c_1^{(2)}$ in the medium $I^{(2)}$. Let us further introduce a Cartesian base $\{\mathbf{e}_1, \mathbf{e}_2\}$, with $\mathbf{e}_1$ running along the crack surface and $\mathbf{e}_2$ perpendicular to $\mathbf{e}_1$ and pointing into the "upper" medium. In this base, every vector can be presented in terms of the corresponding coordinates, so that every position vector $\mathbf{x} = (x_1, x_2)$, every displacement vector $\mathbf{u} = (u_1, u_2)$, etc.

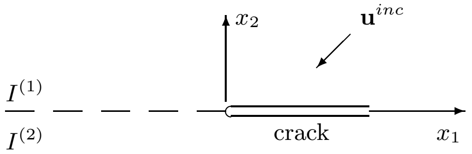


FIG. 2.1. *The problem geometry.*

Let $\mathbf{u}(\mathbf{x}) \exp(-i\omega t)$ be a time harmonic displacement vector in an elastic medium, where t is time, ω is the angular frequency, and the exponential time factor $\exp(-i\omega t)$ is understood but suppressed everywhere below. Using the two dimensional Green's tensors (see Appendix A) and introducing a fictitious bottom medium that has the properties of the upper half space, and a fictitious top medium which has the properties of the bottom half space, the above problem can be recast in the form of a generalized reciprocity relation [11] or the extinction theorem (by analogy with the electromagnetic case—see [12]; the detailed derivation is given in [13]; also see Appendix B). This states that the total displacement for a medium $I^{(j)}$ satisfies the integral equation

$$H[(-1)^{j+1}x_2]u_k^{(j)}(\mathbf{x}) = u_k^{inc(j)}(\mathbf{x}) + (-1)^j \sum_{i=1}^{2} \int_{-\infty}^{-\infty} [\sigma_{2ik}^{G(j)}(x_1 - y_1, x_2)u_i^{(j)}(y_1, 0)$$

$$+ u_{ik}^{G(j)}(x_1 - y_1, x_2)\sigma_{2i}^{(j)}(y_1, 0)]dy_1, \quad j, k = 1, 2, \tag{2.1}$$

where $u^{G(j)}(\mathbf{x} - \mathbf{y})$ and $\sigma^{G(j)}(\mathbf{x} - \mathbf{y})$ are the free space Green's tensor and Green's stress tensor, respectively; $\sigma^{(j)}(\mathbf{x})$ is the stress tensor corresponding to displacement $\mathbf{u}^{(j)}(\mathbf{x})$; $H(x)$ is the Heaviside step function,

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0; \end{cases} \tag{2.2}$$

and in the medium $I^{(j)}$, $j = 1, 2$, the plane wave $\mathbf{u}^{inc(j)}(\mathbf{x})$, which is incident from $I^{(m)}$, is

$$\mathbf{u}^{inc(j)}(\mathbf{x}) = \delta_{mj}\mathbf{d}^{n(m)}e^{-ik_n^{(m)}(p_1^{inc}x_1 - (-1)^m p_2^{inc}x_2)}, \tag{2.3}$$

with δ_{mj}—the Kronecker delta, $\mathbf{p}^{inc} = (p_1^{inc}, p_2^{inc})$—the incoming unit wave vector with $p_2^{inc} > 0$, and $k_n^{(m)} = \omega/c_n^{(m)}$—a wave number (see, e.g., [11]). The longitudinal displacement unit vector $\mathbf{d}^{1(m)}$ is

$$\mathbf{d}^{1(m)} = (-p_1^{inc}, (-1)^m p_2^{inc}), \tag{2.4}$$

and when the motions are transverse the displacement unit vector $\mathbf{d}^{2(m)}$ is

$$\mathbf{d}^{2(m)} = ((-1)^{m+1}p_2^{inc}, -p_1^{inc}). \tag{2.5}$$

To complete the problem statement we require that on the contact boundary, $\{(x_1, x_2) : x_2 = 0,\ x_1 < 0\}$, the displacement and normal stress components be continuous,

$$\begin{aligned} &u_i^{(1)}(x_1, 0) = u_i^{(2)}(x_1, 0), \\ &\sigma_{2i}^{(1)}(x_1, 0) = \sigma_{2i}^{(2)}(x_1, 0), \quad x_1 < 0, \quad i = 1, 2; \end{aligned} \tag{2.6}$$

on the crack $\{(x_1, x_2) : x_2 = 0,\ x_1 > 0\}$ the normal stress components be zero,

$$\sigma_{2i}^{(1)}(x_1, 0) = \sigma_{2i}^{(2)}(x_1, 0) = 0, \quad x_1 > 0, \quad i = 1, 2; \tag{2.7}$$

at infinity, the radiation conditions be satisfied in the form of the limiting absorption principle; and at the crack tip, the mean energy of the diffracted field be bounded. In combination with (2.1), the last condition amounts to requiring that we have

$$\sigma \sim O(r^{-1/2 \pm i\nu_0}), \tag{2.8}$$

with $\nu_0 > 0$ a real bimaterial constant (see, e.g., [14]) and r the distance to the origin $r = \sqrt{x_1^2 + x_2^2}$. The condition (2.8) suggests the oscillatory motions near the crack tip, which is nonphysical. However, this region is often extremely small and in most cases can be ignored. (For further discussion, see Williams [15], Erdogan [16, 17], and Rice and Sih [18]; the related static case has been considered in [19]). Below, if not used as a subscript, $i = \sqrt{-1}$.

3. The functional equations for the Fourier transforms of displacements and stresses in the crack plane. For any field $\varphi(y_1)$ let us define the decomposition

$$\varphi(y_1) = \varphi^+(y_1) + \varphi^-(y_1), \tag{3.1}$$

where the superscripts + and − denote functions that vanish for the negative and positive values of y_1, respectively. Let us use these fields to introduce new "half unknowns," components of the four dimensional vector $\mathbf{v}^\pm(y_1)$ given by

$$\begin{aligned} &v_i^-(y_1) = u_i^{(1)}(y_1, 0) = u_i^{(2)}(y_1, 0), \qquad y_1 < 0,\\ &v_{i+2}^-(y_1) = -\frac{i}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(1)}(y_1, 0) = -\frac{i}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(2)}(y_1, 0), \qquad y_1 < 0,\\ &v_i^+(y_1) = u_i^{(1)}(y_1, 0), \qquad y_1 > 0,\\ &v_{i+2}^+(y_1) = u_i^{(2)}(y_1, 0), \qquad y_1 > 0,\ i = 1, 2. \end{aligned} \tag{3.2}$$

Let us make use of operators of dilatation (or divergence) and rotation (or curl) and denote the dilatation of any tensor $\phi_{ik}^{(j)}(x_1, x_2)$ by the superscript ^{1} and the rotation by the superscript ^{2}, so that we can write

$$\phi_i^{(j)1}(\mathbf{x}) = [\phi_{i1}^{(j)}(\mathbf{x})]_{,1} + [\phi_{i2}^{(j)}(\mathbf{x})]_{,2}\,, \quad \phi_i^{(j)2}(\mathbf{x}) = [\phi_{i2}^{(j)}(\mathbf{x})]_{,1} - [\phi_{i1}^{(j)}(\mathbf{x})]_{,2}\,. \tag{3.3}$$

Applying the dilatation ($l = 1$) and rotation ($l = 2$) to (2.1) in the half plane where the argument of the Heaviside function is negative ($(-1)^j x_2 > 0$) and using the boundary conditions (2.6) and (2.7), the extinction theorem can be rewritten as

$$\begin{aligned} &ik_n^{(m)}\delta_{mj}\delta_{nl}e^{-ik_n^{(m)}(x_1p_1^{inc}+(-1)^{m+1}p_2^{inc}x_2)}\\ &\quad + (-1)^j\int_0^\infty [\sigma_1^{G(j)l}(x_1 - y_1, x_2)v_{2j-1}^+(y_1) + \sigma_2^{G(j)l}(x_1 - y_1, x_2)v_{2j}^+(y_1)]dy_1\\ &\quad + (-1)^j\int_{-\infty}^0 \{\sigma_1^{G(j)l}(x_1 - y_1, x_2)v_1^-(y_1) + \sigma_2^{G(j)l}(x_1 - y_1, x_2)v_2^-(y_1)\\ &\qquad\qquad + ik_1^{(1)}\mu^{(1)}[u_1^{G(j)l}(x_1 - y_1, x_2)v_3^-(y_1)\\ &\qquad\qquad + u_2^{G(j)l}(x_1 - y_1, x_2)v_4^-(y_1)]\}dy_1 = 0, \quad j, l = 1, 2, \end{aligned} \tag{3.4}$$

or in the matrix form as

$$\begin{aligned} &\int_0^\infty A^+(x_1 - y_1, x_2)\mathbf{v}^+(y_1)dy_1 + \int_{-\infty}^0 A^-(x_1 - y_1, x_2)\mathbf{v}^-(y_1)dy_1\\ &\qquad = -ik_n^{(m)}\mathbf{U}^{inc}e^{ik_n^{(m)}\mathbf{x}\cdot\mathbf{d}^{1(m)}}, \end{aligned} \tag{3.5}$$

where $\mathbf{d}^{1(m)}$ is the displacement unit vector of a longitudinal wave defined in (2.4), $\mathbf{U}^{inc}$ is the four dimensional vector

$$\mathbf{U}^{inc} = [\delta_{1m}\delta_{1n},\ \delta_{1m}\delta_{2n},\ \delta_{2m}\delta_{1n},\ \delta_{2m}\delta_{2n}]^T, \tag{3.6}$$

and 4×4 matrices $A^+(\mathbf{x})$ and $A^-(\mathbf{x})$ are

$$A^+(\mathbf{x}) = \begin{pmatrix} -\sigma_1^{G(1)1}(\mathbf{x}) & -\sigma_2^{G(1)1}(\mathbf{x}) & 0 & 0 \\ -\sigma_1^{G(1)2}(\mathbf{x}) & -\sigma_2^{G(1)2}(\mathbf{x}) & 0 & 0 \\ 0 & 0 & \sigma_1^{G(2)1}(\mathbf{x}) & \sigma_2^{G(2)1}(\mathbf{x}) \\ 0 & 0 & \sigma_1^{G(2)2}(\mathbf{x}) & \sigma_2^{G(2)2}(\mathbf{x}) \end{pmatrix},$$

$$A^-(\mathbf{x}) = \begin{pmatrix} -\sigma_1^{G(1)1}(\mathbf{x}) & -\sigma_2^{G(1)1}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_1^{G(1)1}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_2^{G(1)1}(\mathbf{x}) \\ -\sigma_1^{G(1)2}(\mathbf{x}) & -\sigma_2^{G(1)2}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_1^{G(1)2}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_2^{G(1)2}(\mathbf{x}) \\ \sigma_1^{G(2)1}(\mathbf{x}) & \sigma_2^{G(2)1}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_1^{G(2)1}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_2^{G(2)1}(\mathbf{x}) \\ \sigma_1^{G(2)2}(\mathbf{x}) & \sigma_2^{G(2)2}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_1^{G(2)2}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_2^{G(2)2}(\mathbf{x}) \end{pmatrix}, \tag{3.7}$$

with dilatations and rotations $u_i^{G(j)l}(\mathbf{x})$ and $\sigma_i^{G(j)l}(\mathbf{x})$ given in Appendix A.

Everywhere below, let the hat $\widehat{}$ denote the Fourier transform with respect to the nondimensionalized variable $k_1^{(1)}y_1$, so that for any function $\varphi(y_1)$ we have

$$\widehat{\varphi}(\xi) = k_1^{(1)} \int_{-\infty}^{\infty} \varphi(y_1) e^{ik_1^{(1)}\xi y_1} dy_1. \tag{3.8}$$

Let us then take the Fourier transform of (3.5) and evaluate the result on the boundary. For this purpose, let us multiply (3.5) by $\exp(ik_1^{(1)}\xi x_1)$, integrate it over x_1, and set $x_2 = 0$. Applying the convolution theorem, we obtain

$$\begin{aligned} &\frac{1}{k_1^{(1)}} \left[\int_{-\infty}^{\infty} A^+(\mathbf{x}) e^{ik_1^{(1)}\xi x_1} dx_1 \right] \widehat{\mathbf{v}}^+(\xi) \\ &\quad + \frac{1}{k_1^{(1)}} \left[\int_{-\infty}^{\infty} A^-(\mathbf{x}) e^{ik_1^{(1)}\xi x_1} dx_1 \right] \widehat{\mathbf{v}}^-(\xi) = p(\xi)\mathbf{U}^{inc}, \end{aligned} \tag{3.9}$$

where, using the notation $\xi^{inc} = \kappa_n^{(m)} p_1$, $\kappa_n^{(m)} = c_1^{(1)}/c_n^{(m)}$, we have

$$p(\xi) = 2\pi\ i\delta(\xi - \xi^{inc}) = \lim_{\epsilon \to 0} \left(-\frac{1}{\xi - \xi^{inc} + i\epsilon} + \frac{1}{\xi - \xi^{inc} - i\epsilon} \right). \tag{3.10}$$

Multiplying the vector equation in (3.9) by $-2i$ and the first and the third scalar equations there by $[\kappa_2^{(1)}]^2$ and $[\kappa_2^{(2)}/\kappa_1^{(2)}]^2$, respectively, we obtain the system of four scalar functional equations

$$\widehat{A}^+(\xi)\widehat{\mathbf{v}}^+(\xi) + \widehat{A}^-(\xi)\widehat{\mathbf{v}}^-(\xi) = p(\xi)\mathbf{v}^{inc}, \tag{3.11}$$

where the vector $\mathbf{v}^{inc}$ is such that

$$\mathbf{v}^{inc} = 2i \left[(\kappa^{(1)})^2 \delta_{1m}\delta_{1n},\ \delta_{1m}\delta_{2n},\ (\kappa^{(2)})^2 \delta_{2m}\delta_{1n},\ \delta_{2m}\delta_{2n} \right]^T \tag{3.12}$$

with $\kappa^{(j)} = c_1^{(j)}/c_2^{(j)}$; and matrices $\widehat{A}^+(\xi)$ and $\widehat{A}^-(\xi)$ in (3.11) are given by

$$\widehat{A}^+(\xi) = \begin{pmatrix} 2\xi & \frac{a_1(\xi)}{\gamma_1^{(1)}(\xi)} & 0 & 0 \\ -\frac{a_1(\xi)}{\gamma_2^{(1)}(\xi)} & 2\xi & 0 & 0 \\ 0 & 0 & 2\xi & -\frac{a_2(\xi)}{\gamma_1^{(2)}(\xi)} \\ 0 & 0 & \frac{a_2(\xi)}{\gamma_2^{(2)}(\xi)} & 2\xi \end{pmatrix}, \tag{3.13}$$

$$\widehat{A}^-(\xi) = \begin{pmatrix} 2\xi & \frac{a_1(\xi)}{\gamma_1^{(1)}(\xi)} & -\frac{\xi}{\gamma_1^{(1)}(\xi)} & -1 \\ -\frac{a_1(\xi)}{\gamma_2^{(1)}(\xi)} & 2\xi & 1 & -\frac{\xi}{\gamma_2^{(1)}(\xi)} \\ 2\xi & -\frac{a_2(\xi)}{\gamma_1^{(2)}(\xi)} & \frac{\mu\,\xi}{\gamma_1^{(2)}(\xi)} & -\mu \\ \frac{a_2(\xi)}{\gamma_2^{(2)}(\xi)} & 2\xi & \mu & \frac{\mu\,\xi}{\gamma_2^{(2)}(\xi)} \end{pmatrix}, \tag{3.14}$$

with $a_j(\xi) = (\kappa_2^{(j)})^2 - 2\xi^2$, $\gamma_i^{(j)}(\xi) = \sqrt{[\kappa_i^{(j)}]^2 - \xi^2}$, and $\mu = \mu^{(1)}/\mu^{(2)}$.

The dilatation and rotation of the Green's tensor and Green's stress tensor in (3.7) can be expressed in terms of $H_0(k_i^{(j)}r)$, the Hankel function of the first kind of zero order and its derivatives (see Appendix A). Therefore, the matrices $\widehat{A}^{\pm}(\xi)$ have been evaluated with the help of the following identity:

$$\int_{-\infty}^{\infty} H_0(k_i^{(j)}r)e^{ik_1^{(1)}\xi x_1}dx_1 = \frac{2}{k_1^{(1)}}\frac{1}{\gamma_i^{(j)}}e^{ik_1^{(1)}\gamma_i^{(j)}|x_2|} \tag{3.15}$$

(see, e.g., [20]). Note that (3.11) involves singularities, which are described in Appendix C. Using the definition of plus and minus functions together with (3.8), it is easy to check that $\widehat{\mathbf{v}}^+(\xi)$ and $\widehat{\mathbf{v}}^-(\xi)$ are analytic in the upper and lower halves, respectively, of the complex ξ-plane.

Let us cast (3.11) in another form through multiplying by the matrix $[\widehat{A}^-(\xi)]^{-1}$. Then we obtain the vector functional equation

$$\widehat{\mathbf{v}}^-(\xi) + B^+(\xi)\widehat{\mathbf{v}}^+(\xi) = p(\xi)\mathbf{T}^{inc}, \tag{3.16}$$

where the vector on the right-hand side is

$$\mathbf{T}^{inc} = [\widehat{A}^-(\xi^{inc})]^{-1}\mathbf{v}^{inc}; \tag{3.17}$$

the matrix $B^+(\xi) = [\widehat{A}^-(\xi)]^{-1}\widehat{A}^+(\xi)$ is given by

(3.18)

$$B^+(\xi) = \frac{\mu}{S(\xi)} \cdot$$

$$\begin{pmatrix} b_{21}b_{12} - g_1g_2 + \mu R_1h_2 & -(b_{21}g_1 + b_{11}g_2) & b_{11}b_{22} - g_1g_2 + \frac{R_2h_1}{\mu} & b_{21}g_1 + b_{11}g_2 \\ b_{22}g_1 + b_{12}g_2 & b_{22}b_{11} - g_1g_2 + \mu R_1h_2 & -(b_{22}g_1 + b_{12}g_2) & b_{12}b_{21} - g_1g_2 + \frac{R_2h_1}{\mu} \\ -\frac{d_2}{\mu} & \frac{e}{\mu} & \frac{d_2}{\mu} & -\frac{e}{\mu} \\ -\frac{e}{\mu} & -\frac{d_1}{\mu} & \frac{e}{\mu} & \frac{d_1}{\mu} \end{pmatrix},$$

with the Rayleigh functions $R_j(\xi)$, b_{ij}, g_j, $i, j = 1, 2$, and the Stoneley function $S(\xi)$ defined in Appendix C; and

$$\begin{aligned} d_j(\xi) &= R_2(\xi)b_{1j}(\xi) + \mu R_1(\xi)b_{2j}(\xi), \\ e(\xi) &= R_2(\xi)g_1(\xi) - \mu R_1(\xi)g_2(\xi), \quad j = 1, 2. \end{aligned} \tag{3.19}$$

In the case under consideration, the equation $|\widehat{A}^{-}(\xi)| = 0$ has no solutions (see Appendix C), and therefore the matrix $B^{+}(\xi)$ is bounded. Equation (3.16) is a vector functional equation, which has no known analytical solution and is difficult to solve numerically.

4. The functional equations for nonsingular unknowns. Let us reformulate (3.16) as a vector functional equation for nonsingular components of the displacements and stresses in the crack plane. We start by introducing two new unknown vector functions, the crack opening displacement (COD) $\mathbf{u}^{COD}(y_1)$ and $\mathbf{t}(y_1)$, as

$$\begin{aligned} \mathbf{u}^{COD}(y_1) &= \mathbf{u}^{(1)}(y_1, 0) - \mathbf{u}^{(2)}(y_1, 0), \\ \mathbf{t}(y_1) &= \frac{1}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(1)}(y_1, 0). \end{aligned} \tag{4.1}$$

Then using the definition (3.2), their Fourier transforms $\widehat{\mathbf{u}}^{COD}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ may be expressed in terms of components of the old half unknowns $\mathbf{v}^{\pm}(\xi)$ as

$$\begin{aligned} \widehat{u}_l^{COD}(\xi) &= \widehat{v}_l^{+}(\xi) - \widehat{v}_{2+l}^{+}(\xi), \\ \widehat{t}_l(\xi) &= i\widehat{v}_{l+2}^{-}(\xi), \quad l = 1, 2. \end{aligned} \tag{4.2}$$

The last two equations in (3.16) can be rewritten in terms of the new unknowns $\widehat{\mathbf{u}}^{COD}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ as

$$-i\widehat{\mathbf{t}}(\xi) + \tilde{B}^{+}(\xi)\widehat{\mathbf{u}}^{COD}(\xi) = p(\xi)\tilde{\mathbf{T}}^{inc}, \tag{4.3}$$

where $p(\xi)$ is defined in (3.10); the matrix $\tilde{B}^{+}(\xi)$ and vector $\tilde{\mathbf{T}}^{inc}$ are bounded and are

$$\tilde{B}^{+}(\xi) = \frac{1}{S(\xi)}\begin{pmatrix} -d_2(\xi) & e(\xi) \\ -e(\xi) & -d_1(\xi) \end{pmatrix}, \qquad \tilde{\mathbf{T}}^{inc} = (T_3^{inc}, T_4^{inc})^T. \tag{4.4}$$

The right-hand side of the above equation has two poles, $\xi^{inc} + i0$ and $\xi^{inc} - i0$, which lie infinitely close to each other. Let us isolate one of them by introducing

$$\widehat{\mathbf{s}}(\xi) = -i\widehat{\mathbf{t}}(\xi) - \frac{1}{\xi - \xi^{inc} - i0}\tilde{\mathbf{T}}^{inc}. \tag{4.5}$$

Then the second pole $\xi^{inc} - i0$ can be shifted back to the real axis. For this reason, everywhere below we use ξ^{inc} to mean $\xi^{inc} - i0$. Using (4.5), the functional equation (4.3) can be rewritten as

$$\widehat{\mathbf{s}}(\xi) + \tilde{B}^{+}(\xi)\widehat{\mathbf{u}}^{COD}(\xi) = -\frac{1}{\xi - \xi^{inc}}\tilde{\mathbf{T}}^{inc}. \tag{4.6}$$

The function $\widehat{\mathbf{u}}^{COD}(\xi)$ is analytic in the upper half plane. Therefore, its domain contains both Rayleigh poles $-\xi^{R(j)}$, $j = 1, 2$, the incident pole $\xi = \xi^{inc}$, and also

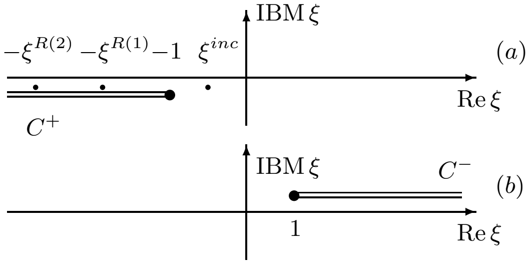


FIG. 4.1. *The poles and branch cuts of* (a) $\widehat{\mathbf{u}}^{COD}(\xi)$ *and* (b) $\widehat{\mathbf{s}}(\xi)$.

the branch cut C^+ (see Figure 4.1(a) and Appendix C). This means that we can decompose $\widehat{\mathbf{u}}^{COD}(\xi)$ as follows:

$$\widehat{\mathbf{u}}^{COD}(\xi) = \frac{G(\xi^{inc})}{G(\xi)}\widehat{\mathbf{w}}^{ge+}(\xi) + \frac{G(\xi)}{[\xi+\xi^{R(1)}][\xi+\xi^{R(2)}]}\widehat{\mathbf{w}}^{+}(\xi), \tag{4.7}$$

where $G(\xi)$ is defined in Appendix D, $\widehat{\mathbf{w}}^{ge+}(\xi)$ has the pole at $\xi = \xi^{inc}$, and $\widehat{\mathbf{w}}^{+}(\xi)$ is a new unknown vector describing the edge diffracted body and surface waves. Note that as $\xi \to \infty$, the function $G(\xi) \sim \xi$, and therefore $\widehat{\mathbf{w}}^{+}(\xi)$ behaves as $\xi\widehat{\mathbf{u}}^{COD}(\xi)$.

The function $\widehat{\mathbf{w}}^{ge+}(\xi)$ can be chosen to be

$$\begin{aligned}\widehat{\mathbf{w}}^{ge+}(\xi) = &-\frac{1}{\xi-\xi^{inc}}\tilde{B}^{-}(\xi^{inc})\tilde{\mathbf{T}}^{inc}\\ &+\frac{1}{\xi-\xi^{inc}}\frac{1}{G(\xi^{inc})}\sum_{k=1}^{4}F_k^{+}(\xi,\xi^{inc})\Delta_k^{+}[G(\xi)\tilde{B}^{-}(\xi)]\big|_{\xi=\xi^{inc}}\tilde{\mathbf{T}}^{inc},\end{aligned} \tag{4.8}$$

where the auxiliary functions $F_k^{+}(\xi,\zeta)$, $k = 1,\dots,4$, are defined in Appendix D; we have

$$\tilde{B}^{-}(\xi) = \frac{1}{R_1(\xi)R_2(\xi)}\begin{pmatrix} -d_1(\xi) & -e(\xi) \\ e(\xi) & -d_2(\xi)\end{pmatrix}, \tag{4.9}$$

and for any function $\Phi(\xi)$ we denote by $\Delta_k^{+}\Phi(\xi)$ the jump of $\Phi(\xi)$ over the branch cut C_k^{+} defined in (D.4) of Appendix D.

Analogously, the function $\widehat{\mathbf{s}}(\xi)$ is analytic in the lower half plane, and we can write the decomposition

$$\widehat{\mathbf{s}}(\xi) = \widehat{\mathbf{w}}^{ge-}(\xi) + \widehat{\mathbf{w}}^{-}(\xi), \tag{4.10}$$

where $\widehat{\mathbf{w}}^{ge-}(\xi)$ is introduced to cancel the extraneous incident pole appearing on the negative side of the branch cut C^- (see Figure 4.1(b)) and $\widehat{\mathbf{w}}^{-}(\xi)$ is a new unknown vector. The function $\widehat{\mathbf{w}}^{ge-}(\xi)$ can be chosen to be

$$\widehat{\mathbf{w}}^{ge-}(\xi) = -\frac{1}{\xi-\xi^{inc}}\sum_{k=1}^{4}F_k^{-}(\xi,\xi^{inc})\Delta_k^{-}[\tilde{B}^{+}(\xi)]\big|_{\xi=\xi^{inc}}\tilde{B}^{-}(\xi^{inc})\tilde{\mathbf{T}}^{inc}, \tag{4.11}$$

where the auxiliary functions $F_k^{-}(\xi,\zeta)$ are defined in Appendix D and $\Delta_k^{-}\Phi(\xi)$ denotes a jump of a function $\Phi(\xi)$ over the branch cut C_k^{-} defined in (D.4). It can be shown that whether or not $\xi = \xi^{inc}$ lies close to the branch cut C^-, we can always use the definition (4.11).

Let us substitute decompositions (4.7) and (4.10) into the functional equation (4.6) to obtain a functional equation for the new unknowns $\widehat{\mathbf{w}}^+(\xi)$ and $\widehat{\mathbf{w}}^-(\xi)$,

$$\widehat{\mathbf{w}}^-(\xi) + M^+(\xi)\widehat{\mathbf{w}}^+(\xi) = \mathbf{V}^{inc-}(\xi), \tag{4.12}$$

where the matrix $M^+(\xi)$ is given by

$$M^+(\xi) = \frac{G(\xi)}{(\xi^{R(1)} + \xi)(\xi^{R(2)} + \xi)}\tilde{B}^+(\xi) \tag{4.13}$$

and the vector $\mathbf{V}^{inc-}(\xi)$ on the right-hand side of (4.12) can be written as

$$\mathbf{V}^{inc-}(\xi) = -\frac{1}{\xi - \xi^{inc}}\tilde{\mathbf{T}}^{inc} - \frac{G(\xi^{inc})}{G(\xi)}\tilde{B}^+(\xi)\widehat{\mathbf{w}}^{ge+}(\xi) - \widehat{\mathbf{w}}^{ge-}(\xi). \tag{4.14}$$

Multiplying (4.12) by the inverse matrix $[M^+(\xi)]^{-1}$, we obtain the functional equation

$$\widehat{\mathbf{w}}^+(\xi) + M^-(\xi)\widehat{\mathbf{w}}^-(\xi) = \mathbf{V}^{inc+}(\xi), \tag{4.15}$$

where $M^-(\xi) = [M^+(\xi)]^{-1}$, so that we have

$$M^-(\xi) = \frac{(\xi^{R(1)} + \xi)(\xi^{R(2)} + \xi)}{G(\xi)R_1(\xi)R_2(\xi)} \begin{pmatrix} -d_1(\xi) & -e(\xi) \\ e(\xi) & -d_2(\xi) \end{pmatrix}, \tag{4.16}$$

and the vector $\mathbf{V}^{inc+}(\xi)$ is

$$\mathbf{V}^{inc+}(\xi) = M^-(\xi)\mathbf{V}^{inc-}(\xi). \tag{4.17}$$

The functional vector equations (4.12) and (4.15) form a system of four functional equations in four unknowns.

5. The system of integral equations.

Since the vector $\widehat{\mathbf{w}}^+(\xi)$ introduced in the decomposition (4.7) has no poles and vanishes at infinity, it can be represented as a Hilbert transform

$$\widehat{\mathbf{w}}^+(\xi) = -\frac{1}{2\pi i}\int_1^\infty \frac{\Delta\mathbf{w}^+(\xi')}{\xi' + \xi}d\xi', \tag{5.1}$$

where $\Delta\mathbf{w}^+(\xi)$ is a jump of $\widehat{\mathbf{w}}^+(\xi)$ over the branch cut C^+ (see Figure 4.1(a)); i.e., we have

$$\Delta\mathbf{w}^+(\xi) = \widehat{\mathbf{w}}^+(-\xi + i0) - \widehat{\mathbf{w}}^+(-\xi - i0), \quad \xi > 1. \tag{5.2}$$

Also, the vector $\widehat{\mathbf{w}}^-(\xi)$ introduced in (4.10) can be represented as

$$\widehat{\mathbf{w}}^-(\xi) = -\frac{1}{2\pi i}\int_1^\infty \frac{\Delta\mathbf{w}^-(\xi')}{\xi' - \xi}d\xi', \tag{5.3}$$

where, for $\xi \in C^-$, $\Delta\mathbf{w}^-(\xi)$ is a jump of $\widehat{\mathbf{w}}^-(\xi)$ over the branch cut C^-(see Figure 4.1(b)); i.e., we have

$$\Delta\mathbf{w}^-(\xi) = \widehat{\mathbf{w}}^-(\xi - i0) - \widehat{\mathbf{w}}^-(\xi + i0), \quad \xi > 1. \tag{5.4}$$

Note that $\widehat{\mathbf{w}}^+(\xi)$ and $\widehat{\mathbf{w}}^-(\xi)$ have the same branches as $\widehat{\mathbf{v}}^+(\xi)$ and $\widehat{\mathbf{v}}^-(\xi)$, respectively (see, e.g., [9]).

Let us now substitute (5.1) into (4.15), (5.3) into (4.12), and calculate the jumps in the resulting equations across the branch cuts C^+ and C^-, respectively. Then for $\xi > 1$ we have

$$\begin{aligned} &\Delta\mathbf{w}^+(\xi) + \Delta M^-(-\xi)\widehat{\mathbf{w}}^-(-\xi) = \Delta\mathbf{V}^{inc+}(\xi), \\ &\Delta\mathbf{w}^-(\xi) + \Delta M^+(\xi)\widehat{\mathbf{w}}^+(\xi) = \Delta\mathbf{V}^{inc-}(\xi), \end{aligned} \tag{5.5}$$

where Δ denotes the jump over the cut, $\xi \in C^-$, and we use the notation

$$\Delta M^\pm(\pm\xi) = 2i\,\mathrm{IBM}\,M^\pm(\pm\xi), \quad \xi \geq 1. \tag{5.6}$$

In view of (3.10), (4.16), (5.2), (5.4), and the definitions of the functions $F_k^\pm(\xi, a)$, $k = 1, \dots, 4$ (see Appendix D), the apparent poles $\xi = \xi^{R(j)}$ and $\xi = \pm\xi^{inc}$ in (5.5) are in fact absent (their residues are zero). Decompositions (4.8) and (4.11) have been chosen in order to achieve this regularization.

Equation (5.5) can be rewritten as the system of coupled integral equations

$$\begin{aligned} &\Delta\mathbf{w}^+(\xi) - \frac{1}{\pi}\mathrm{IBM}\,[M^-(-\xi)] \int_1^\infty \frac{\Delta\mathbf{w}^-(\xi')}{\xi+\xi'}d\xi' = \Delta\mathbf{V}^{inc+}(\xi), \\ &\Delta\mathbf{w}^-(\xi) - \frac{1}{\pi}\mathrm{IBM}\,[M^+(\xi)] \int_1^\infty \frac{\Delta\mathbf{w}^+(\xi')}{\xi+\xi'}d\xi' = \Delta\mathbf{V}^{inc-}(\xi). \end{aligned} \tag{5.7}$$

6. The numerical scheme. As mentioned above, we seek the solution of (5.7) that allows the tip condition (2.8) to be satisfied, that is, exhibit the asymptotic behavior

$$\Delta\mathbf{w}(\xi) \to \xi^{-1/2}[\mathbf{D}^+\cos(\nu_0 \ln\xi) + \mathbf{D}^-\sin(\nu_0\ln\xi)], \quad \xi \to \infty, \tag{6.1}$$

where ν_0 is a bimaterial constant; $\mathbf{D}^\pm$ are unknown four dimensional constant vectors, and $\Delta\mathbf{w}(\xi) = (\Delta w_1^+(\xi), \Delta w_2^+(\xi), \Delta w_1^-(\xi), \Delta w_2^-(\xi))$ is the unknown four dimensional vector function.

Let us rewrite the system of (5.7) as one vector integral equation,

$$\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\int_1^\infty \frac{\Delta\mathbf{w}(\xi')}{\xi'+\xi}d\xi' = \Delta\mathbf{V}^{inc}(\xi), \tag{6.2}$$

where we use the notation

$$\Delta\mathbf{V}^{inc}(\xi) = (\Delta V_1{}^{inc+}(\xi), \Delta V_2{}^{inc+}(\xi), \Delta V_1{}^{inc-}(\xi), \Delta V_2{}^{inc-}(\xi))^T,$$

with $\Delta V_i{}^{inc\pm}(\xi),\ i = 1, 2$, being the right-hand side of (5.5) and $M(\xi)$ being the 4×4 matrix that can be written as

$$M(\xi) = \begin{pmatrix} 0_2 & \mathrm{IBM}\,[M^-(-\xi)] \\ \mathrm{IBM}\,[M^+(\xi)] & 0_2 \end{pmatrix}. \tag{6.3}$$

Then there exists a constant L_2 such that for $\xi \geq L_2$ the solution $\Delta\mathbf{w}(\xi)$ exhibits the behavior (6.1) and we can rewrite (6.2) as

$$\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\int_1^{L_2} \frac{\Delta\mathbf{w}(\xi')}{\xi'+\xi}d\xi' - \frac{1}{\pi}M(\xi)[\mathbf{D}^+I^+(\xi) + \mathbf{D}^-I^-(\xi)] = \Delta\mathbf{V}^{inc}(\xi), \tag{6.4}$$

where we use the notation

$$\begin{aligned} I^{+}(\xi) &= \frac{1}{2}\left(\int_{L_2}^{\infty} \frac{x^{-1/2+i\nu_0}}{x+\xi}dx + \int_{L_2}^{\infty} \frac{x^{-1/2-i\nu_0}}{x+\xi}dx\right), \\ I^{-}(\xi) &= \frac{1}{2i}\left(\int_{L_2}^{\infty} \frac{x^{-1/2+i\nu_0}}{x+\xi}dx - \int_{L_2}^{\infty} \frac{x^{-1/2-i\nu_0}}{x+\xi}dx\right). \end{aligned} \tag{6.5}$$

The integral in (6.4) can be approximated using N collocation points $\xi_i,\ i = 1, N$, and weights $W_i,\ i = 1, N$, so that (6.4) may be approximated by

$$\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\sum_{i=1}^{N}\frac{\Delta\mathbf{w}(\xi_i)}{\xi_i+\xi}W_i - \frac{1}{\pi}M(\xi)[\mathbf{D}^{+}I^{+}(\xi) + \mathbf{D}^{-}I^{-}(\xi)] = \Delta\mathbf{V}^{inc}(\xi), \tag{6.6}$$

which at $\xi = \xi_j,\ j = 1, N$, becomes

$$\begin{aligned} \Delta\mathbf{w}(\xi_j) - \frac{1}{\pi}M(\xi_j)\sum_{i=1}^{N}\frac{\Delta\mathbf{w}(\xi_i)}{\xi_i+\xi_j}W_i - \frac{1}{\pi}M(\xi_j)[\mathbf{D}^{+}I^{+}(\xi_j) + \mathbf{D}^{-}I^{-}(\xi_j)] &= \Delta\mathbf{V}^{inc}(\xi_j), \\ j &= 1, \ldots, N. \end{aligned} \tag{6.7}$$

System (6.7) contains $4N$ equations with $4N+8$ unknowns $\Delta w_i(\xi_j),\ i = 1, 4,\ j = 1, N$, and $D_i^{\pm},\ i = 1, 4$, and is thus underdetermined. Four extra equations are provided by the continuity of the solution at the point $\xi_N = L_2$ and can be written as

$$\Delta\mathbf{w}(\xi_N) = \xi_N^{-1/2}\left[\mathbf{D}^{+}\cos(\nu_0 \ln \xi_N) + \mathbf{D}^{-}\sin(\nu_0 \ln \xi_N)\right]. \tag{6.8}$$

Four more can be supplied by the relation (E.15) (see Appendix E). The resulting algebraic system of $4N + 8$ equations in $4N + 8$ unknowns can be written as

$$M\Delta\mathbf{w} = \mathbf{f}, \tag{6.9}$$

where the $4N + 8$ dimensional vectors are

$$\begin{aligned} \Delta\mathbf{w} &= (\Delta w_1(\xi_1), \Delta w_2(\xi_1), \Delta w_3(\xi_1), \Delta w_4(\xi_1), w_1(\xi_2), \ldots, \\ &\qquad \Delta w_4(\xi_N), D_1^{+}, \ldots, D_4^{+}, D_1^{-}, \ldots, D_4^{-}), \\ \mathbf{f} &= (\Delta V_1^{inc}(\xi_1), \ldots, \Delta V_4^{inc}(\xi_1), \Delta V_1^{inc}(\xi_2), \ldots, \Delta V_4^{inc}(\xi_N), 0, 0, 0, 0, 0, 0, 0, 0), \end{aligned} \tag{6.10}$$

and matrix M may be given in a block form as

$$M = \begin{pmatrix} I_4 - \frac{M(\xi_1)W_1}{2\pi\xi_1} & -\frac{M(\xi_1)W_2}{\pi(\xi_1+\xi_2)} & \cdots & -\frac{M(\xi_1)W_{N-1}}{\pi(\xi_1+\xi_{N-1})} & -\frac{M(\xi_1)W_N}{\pi(\xi_1+\xi_N)} & M^D(\xi_1) \\ -\frac{M(\xi_2)W_1}{\pi(\xi_1+\xi_2)} & I_4 - \frac{M(\xi_2)W_2}{2\pi\xi_2} & \cdots & -\frac{M(\xi_2)W_{N-1}}{\pi(\xi_2+\xi_{N-1})} & -\frac{M(\xi_2)W_N}{\pi(\xi_2+\xi_N)} & M^D(\xi_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{M(\xi_N)W_1}{\pi(\xi_1+\xi_N)} & -\frac{M(\xi_N)W_2}{\pi(\xi_2+\xi_N)} & \cdots & -\frac{M(\xi_N)W_{N-1}}{\pi(\xi_N+\xi_{N-1})} & I_4 - \frac{M(\xi_N)W_N}{2\pi\xi_N} & M^D(\xi_N) \\ 0_4 & 0_4 & \cdots & 0_4 & I_4 & M^{V1} \\ 0_4 & 0_4 & \cdots & 0_4 & 0_4 & M^{V2} \end{pmatrix}, \tag{6.11}$$

where we use the notation

$$\begin{aligned} M^D(\xi) &= -\frac{1}{\pi} M(\xi) J\left(I^+(\xi), I^-(\xi)\right), \\ M^{V1} &= -J\left(\xi_N^{-1/2}\cos(\nu_0 \ln \xi_N), \xi_N^{-1/2}\sin(\nu_0 \ln \xi_N)\xi_N)\right), \end{aligned} \tag{6.12}$$

with matrix $J\left(f_1(\xi), f_2(\xi)\right)$ given by

$$J\left(f_1(\xi), f_2(\xi)\right) = \begin{pmatrix} f_1(\xi) & f_2(\xi) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_1(\xi) & f_2(\xi) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_1(\xi) & f_2(\xi) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & f_1(\xi) & f_2(\xi) \end{pmatrix}, \tag{6.13}$$

and

$$M^{V2} = \begin{pmatrix} I_2 & -\sqrt{\frac{m_\infty^-}{m_\infty^+}} I_2 & 0_2 & 0_2 \\ 0_2 & 0_2 & I_2 & -\sqrt{\frac{m_\infty^-}{m_\infty^+}} I_2 \end{pmatrix}. \tag{6.14}$$

To evaluate integrals $I^{\pm}(\xi)$ in (6.5), we consider two cases:

- Case 1: $\xi \le L_2$. It is easy to see that we can write

$$\frac{x}{x+\xi} = \sum_{k=0}^{\infty} (-1)^k \left(\frac{\xi}{x}\right)^k, \tag{6.15}$$

where the right-hand side is the geometric series with the common ratio $-\xi/x$ such that $|\xi/x| < 1$. Therefore, we have

$$\int_{L_2}^{\infty} \frac{x^{-1/2 \pm i\nu_0}}{x+\xi} dx = \int_{L_2}^{\infty} x^{-3/2 \pm i\nu_0} \sum_{k=0}^{\infty} (-1)^k \left(\frac{\xi}{x}\right)^k dx. \tag{6.16}$$

Exchanging in the right-hand side the order of summation and integration and evaluating the resulting integrals, we obtain

$$\int_{L_2}^{\infty} \frac{x^{-1/2 \pm i\nu_0}}{x+\xi} dx = (L_2)^{-1/2 \pm i\nu_0} \sum_{k=0}^{\infty} \frac{(-1)^k}{0.5 + k \mp i\nu_0} \left(\frac{\xi}{L_2}\right)^k. \tag{6.17}$$

Substituting (6.17) into (6.5) gives us

$$\begin{aligned} I^+(\xi) &= \frac{1}{\sqrt{L_2}} \left\{ S_1\left(\frac{\xi}{L_2}\right) \cos(\nu_0 \ln L_2) - \nu_0 S_2\left(\frac{\xi}{L_2}\right) \sin(\nu_0 \ln L_2) \right\}, \\ I^-(\xi) &= \frac{1}{\sqrt{L_2}} \left\{ S_1\left(\frac{\xi}{L_2}\right) \sin(\nu_0 \ln L_2) + \nu_0 S_2\left(\frac{\xi}{L_2}\right) \cos(\nu_0 \ln L_2) \right\}, \end{aligned} \tag{6.18}$$

where we use the notation

$$S_1(\xi) = \sum_{k=0}^{\infty} \frac{(-1)^k (0.5+k)}{(0.5+k)^2 + \nu_0^2} \xi^k, \qquad S_2(\xi) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(0.5+k)^2 + \nu_0^2} \xi^k. \tag{6.19}$$

Note that, according to the Leibnitz theorem, the alternating series $S_i(\xi/L_2)$, $i=1,2$, converge. The difference between the infinite series and partial sum $S_{iN}(\xi/L_2)$ does not exceed the $(N+1)$th term. This property can be used to establish the number of terms necessary to achieve the required accuracy. The latter can be improved further if instead of a partial sum S_{iN} we use S_{iN} plus one half of the $(N+1)$th term. Then the error in $S_i(\xi/L_2)$, $i=1,2$, is $O(k^{-i-1})$ rather than $O(k^{-i})$.

- Case 2: $\xi > L_2$. Let us rewrite each term in (6.5) as

$$\int_{L_2}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \int_0^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx - \int_0^{L_2} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx. \tag{6.20}$$

The first integral can be evaluated to give

$$\int_0^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \frac{\pi}{\cosh \pi\nu_o}\xi^{-1/2\pm i\nu_0} \tag{6.21}$$

(see, e.g., [20]). The integrand of the second term in (6.20) can be represented as

$$\frac{x^{-1/2\pm i\nu_0}}{x+\xi} = \sum_{k=0}^{\infty} \frac{x^{-1/2+k\pm i\nu_0}}{\xi^{k+1}}. \tag{6.22}$$

Integrating both sides of (6.22) over $\xi \in [0, L_2]$ and using (6.21), the integral (6.20) can be represented as

(6.23)

$$\int_{L_2}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \frac{\pi}{\cosh \pi\nu_o}\xi^{-1/2\pm i\nu_0} - \frac{(L_2)^{0.5\pm i\nu_0}}{\xi}\sum_{k=0}^{\infty} \frac{(-1)^k}{0.5+k\pm i\nu_0}\left(\frac{L_2}{\xi}\right)^k.$$

Therefore, (6.5) becomes

$$\begin{aligned} I^+(\xi) &= -\frac{\sqrt{L_2}}{\xi}\left\{S_1\left(\frac{L_2}{\xi}\right)\cos(\nu_0 \ln L_2) + \nu_0 S_2\left(\frac{L_2}{\xi}\right)\sin(\nu_0 \ln L_2)\right\} \\ &\quad + \frac{\pi}{\sqrt{\xi}\cosh \pi\nu_0}\cos(\nu_0 \ln \xi), \\ I^-(\xi) &= -\frac{\sqrt{L_2}}{\xi}\left\{S_1\left(\frac{L_2}{\xi}\right)\sin(\nu_0 \ln L_2) - \nu_0 S_2\left(\frac{L_2}{\xi}\right)\cos(\nu_0 \ln L_2)\right\} \\ &\quad + \frac{\pi}{\sqrt{\xi}\cosh \pi\nu_0}\sin(\nu_0 \ln \xi). \end{aligned} \tag{6.24}$$

Note that, similarly to the previous case, according to the Leibnitz theorem, the alternating series $S_i(L_2/\xi)$, $i=1,2$, converge, and instead of the partial sum S_{iN}, we can utilize S_{iN} plus one half of the $(N+1)$th term.

The resulting linear system (6.9) can be solved numerically using the LU decomposition subroutines from the LAPACK library. After the unknowns are evaluated at the nodes, system (6.6) can be used to extrapolate the solution for any $\xi \geq 1$. Note that in system (6.9) the integrals $I^{\pm}(\xi)$ are calculated at $\xi \leq L_2$, and therefore when solving this system we utilize only (6.18). When extrapolating, (6.24) is used instead. It is easy to see that, whatever the case, the integrals $I^{\pm}(\xi)$ are real valued.

7. The diffraction coefficients.

7.1. The diffraction coefficients for bulk waves. Let us first define the function

$$E(k_i^{(j)}r) = \sqrt{\frac{i}{8\pi}}\frac{e^{ik_i^{(j)}r}}{\sqrt{k_i^{(j)}r}}e^{-ik_i^{(j)}y_1p_1}, \tag{7.1}$$

where $\mathbf{p} = (\cos\theta, \sin\theta)$ is a unit vector in the direction of the diffracted wave, with the angle θ such that $\cos\theta = x_1/r$ and r is the distance to the origin. Let $\mathbf{p}^{\perp}$ be another unit vector, which is orthogonal to $\mathbf{p}$. It then can be chosen to be $\mathbf{p}^{\perp} = (-p_2, p_1) = (-\sin\theta, \cos\theta)$. It is easy to check that as $k_i^{(j)}r \to \infty$, the leading term in the expansion of the argument gives

$$\begin{aligned} H_0\left(k_i^{(j)}\sqrt{(x_1-y_1)^2+x_2^2}\right) &= H_0\left(k_i^{(j)}\left[r - y_1\cos\theta + O\left(\frac{1}{r}\right)\right]\right) \\ &\approx H_0(k_i^{(j)}r) \approx -4iE(k_i^{(j)}r). \end{aligned} \tag{7.2}$$

In the far field, the Green's tensor and Green's stress can be decomposed as

$$\begin{aligned} u_{ik}^{G(j)}(x_1-y_1,x_2) &= u_{ik}^{G1(j)}(x_1-y_1,x_2) + u_{ik}^{G2(j)}(x_1-y_1,x_2), \\ \sigma_{2ik}^{G(j)}(x_1-y_1,x_2) &= \sigma_{2ik}^{G1(j)}(x_1-y_1,x_2) + \sigma_{2ik}^{G2(j)}(x_1-y_1,x_2), \end{aligned} \tag{7.3}$$

where the superscript $i(j)$ refers to longitudinal ($i = 1$) or transverse ($i = 2$) wave in the medium $I^{(j)}$, and in the far field, as $k_i^{(j)}r \to \infty$, we can use approximations

$$\begin{aligned} u_{ik}^{G1(j)}(x_1-y_1,x_2) &\approx \left[\frac{1}{\mu^{(j)}}(\kappa^{(j)})^{-2}p_iE(k_1^{(j)}r)\right]p_k, \\ \sigma_{2ik}^{G1(j)}(x_1-y_1,x_2) &\approx ik_1^{(j)}\left[(\delta_{i2} - 2(\kappa^{(j)})^{-2}p_1p_i^{\perp})E(k_1^{(j)}r)\right]p_k, \\ u_{ik}^{G2(j)}(x_1-y_1,x_2) &\approx \left[\frac{1}{\mu^{(j)}}p_i^{\perp}E(k_2^{(j)}r)\right]p_k^{\perp}, \\ \sigma_{2ik}^{G2(j)}(x_1-y_1,x_2) &\approx ik_2^{(j)}\left[(p_2p_i^{\perp} + p_1p_i)E(k_2^{(j)}r)\right]p_k^{\perp} \end{aligned} \tag{7.4}$$

(see [11], [21], or Appendix A).

To continue, for the scattered field $\mathbf{u}^{sc(j)}(\mathbf{x})$, the integral representation (2.1) can be rewritten as

$$\begin{aligned} H[(-1)^{j+1}x_2]u_k^{sc(j)}(\mathbf{x}) = (-1)^j\sum_{i=1}^{2}\int_{-\infty}^{\infty}&[\sigma_{2ik}^{G(j)}(x_1-y_1,x_2)u_i^{sc(j)}(y_1,0) \\ &+ u_{ik}^{G(j)}(x_1-y_1,x_2)\sigma_{2i}^{(j)sc}(y_1,0)]dy_1 \end{aligned} \tag{7.5}$$

(see, e.g., [11]). Substituting (7.4) into (7.3) and the result into the version of (2.1) applicable to the scattered field, for $0 < \theta < 2\pi$ we can write

$$H(\pi-\theta)u_k^{tip(j)} \approx D^{1(j)}(\theta)\frac{e^{ik_1^{(j)}r}}{\sqrt{k_1^{(j)}r}}p_k + D^{2(j)}(\theta)\frac{e^{ik_2^{(j)}r}}{\sqrt{k_2^{(j)}r}}p_k^{\perp},$$

where the diffraction coefficients for diffracted body waves can be expressed in terms of the displacement vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, and traction vector $\widehat{\mathbf{t}}(\xi)$ in the following manner:

$$\begin{aligned}
D^{1(1)}(\theta) &= -\sqrt{\frac{-i}{8\pi}}(\kappa^{(1)})^{-2}\Big\{2p_1p_2\widehat{u}_1^{(1)}(-p_1)+[(\kappa^{(1)})^2-2p_1^2]\widehat{u}_2^{(1)}(-p_1)\\
&\qquad\qquad +p_1\widehat{t}_1(-p_1)+p_2\widehat{t}_2(-p_1)\Big\},\\
D^{1(2)}(\theta) &= \sqrt{\frac{-i}{8\pi}}\left[\kappa_1^{(2)}\{2(\kappa^{(2)})^{-2}p_1p_2\widehat{u}_1^{(2)}(-\kappa_1^{(2)}p_1)+[1-2(\kappa^{(2)})^{-2}p_1^2]\widehat{u}_2^{(2)}(-\kappa_1^{(2)}p_1)\}\right.\\
&\qquad \left.+\mu(\kappa^{(2)})^{-2}\{p_1\widehat{t}_1(-\kappa_1^{(2)}p_1)+p_2\widehat{t}_2(-\kappa_1^{(2)}p_1)\}\right],\\
D^{2(1)}(\theta) &= -\sqrt{\frac{-i}{8\pi}}\left[\kappa^{(1)}\{[p_1^2-p_2^2]\widehat{u}_1^{(1)}(-\kappa_2^{(1)}p_1)+2p_1p_2\widehat{u}_2^{(1)}(-\kappa_2^{(1)}p_1)\}\right.\\
&\qquad \left.-p_2\widehat{v}_3^-(-\kappa_2^{(1)}p_1)+p_1\widehat{v}_4^-(-\kappa_2^{(1)}p_1)\right],\\
D^{2(2)}(\theta) &= \sqrt{\frac{-i}{8\pi}}\left[\kappa_2^{(2)}\{[p_1^2-p_2^2]\widehat{u}_1^{(2)}(-\kappa_2^{(2)}p_1)+2p_1p_2\widehat{u}_2^{(2)}(-\kappa_2^{(2)}p_1)\}\right.\\
&\qquad \left.+\mu\{-p_2\widehat{v}_3^-(-\kappa_2^{(2)}p_1)+p_1\widehat{v}_4^-(-\kappa_2^{(2)}p_1)\}\right],
\end{aligned}\tag{7.6}$$

where if $\widehat{\mathbf{s}}(\xi)$ is known, $\widehat{\mathbf{t}}(\xi)$ can be found using (4.5). An additional difficulty is presented by the fact that, for $\xi>1$, $\widehat{\mathbf{s}}(\xi)$ as defined in (4.10) contains a singular integral (5.3). Since in (7.6) the arguments of all $\widehat{\mathbf{t}}$ components are given in the form $\xi=-\kappa_i^{(j)}p_1$, the function that needs evaluating is $\widehat{\mathbf{s}}(-\kappa_i^{(j)}p_1)$. When $p_1\geq 0$ the evaluation can be carried out using (4.10), and when $p_1\leq 0$ the combination of (4.7) and (4.6) should be used instead. To calculate the displacement vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, we note that the system (3.11) can be rewritten in terms of $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, and $\widehat{\mathbf{t}}(\xi)$ as

$$p(\xi)\mathbf{v}^{(j)inc}=\widehat{A}^{(j)+}(\xi)\widehat{\mathbf{u}}^{(j)}(\xi)+\widehat{A}^{(j)-}(\xi)\widehat{\mathbf{t}}(\xi),\quad j=1,2,\tag{7.7}$$

where $\widehat{A}^{(j)\pm}(\xi)$ are 2×2 matrices, which form matrices $\widehat{A}^{\pm}(\xi)$ in (3.11), so that we have

$$\widehat{A}^{+}(\xi)=\left(\begin{array}{cc}\widehat{A}^{(1)+}(\xi) & 0\\ 0 & \widehat{A}^{(2)+}(\xi)\end{array}\right),\qquad A^{-}(\xi)=\left(\begin{array}{cc}\widehat{A}^{(1)+}(\xi) & \widehat{A}^{(1)-}(\xi)\\ \widehat{A}^{(2)+}(\xi) & \widehat{A}^{(2)-}(\xi)\end{array}\right),\tag{7.8}$$

and vectors $\mathbf{v}^{(j)inc}(\xi)$ are

$$\mathbf{v}^{(1)inc}(\xi)=\left(\begin{array}{c}v_1^{inc}(\xi)\\ v_2^{inc}(\xi)\end{array}\right),\quad \mathbf{v}^{(2)inc}(\xi)=\left(\begin{array}{c}v_3^{inc}(\xi)\\ v_4^{inc}(\xi)\end{array}\right).\tag{7.9}$$

Therefore, after $\widehat{\mathbf{t}}(\xi)$ is found, vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, can be calculated using (7.7). Note that since $|\kappa_i^{(j)}p_1|\leq\max\{\kappa_i^{(j)}\}<\xi^{R(j)}$, matrices $\widehat{A}^{(j)+}(-\kappa_i^{(j)}p_1)$ are regular.

7.2. The Rayleigh diffraction coefficients. On each of the traction-free surfaces $(x_2=0, x_1>0)$ the Rayleigh wave can be defined as

$$\mathbf{u}^{R(j)}(y_1,0)=D^{(j)R}\mathbf{v}^{(j)R}e^{ik^{R(j)}y_1},\qquad y_1>0,\tag{7.10}$$

where the so-called Rayleigh diffraction coefficient $D^{(j)R}$ is its amplitude, $k^{R(j)} = k_1^{(1)}\xi^{R(j)}$, and the unit vector $\mathbf{v}^{(j)R}$ is a nonzero solution of the homogeneous equation

$$\widehat{A}^{+}(-\xi^{R(j)})\mathbf{v}^{(j)R} = \mathbf{0}. \tag{7.11}$$

It can be shown that the unit vector $\mathbf{v}^{(j)R}$ is

$$\mathbf{v}^{(j)R} = \left(-\frac{2\xi^{R(j)}\gamma_2^{(j)}(\xi^{R(j)})}{[\kappa_2^{(j)}]^2}, (-1)^{j+1}\frac{a_j(\xi^{R(j)})}{[\kappa_2^{(j)}]^2}\right)^T. \tag{7.12}$$

Applying the Fourier transform to (7.10), we obtain

$$\widehat{\mathbf{u}}^{R(j)}(\xi) = \frac{iD^{(j)R}\mathbf{v}^{(j)R}}{\xi + \xi^{R(j)}}. \tag{7.13}$$

Multiplying (7.7) by $[\widehat{A}^{(j)+}(\xi)]^{-1}$ gives us

$$p(\xi)[\widehat{A}^{(j)+}(\xi)]^{-1}\mathbf{v}^{(j)inc} = \widehat{\mathbf{u}}^{(j)}(\xi) + \frac{1}{R_j}\widehat{B}^{(j)-}(\xi)\widehat{\mathbf{t}}(\xi), \quad j = 1, 2, \tag{7.14}$$

with a finite matrix

$$\widehat{B}^{(j)-}(\xi) = \mu_j \begin{pmatrix} (-1)^j b_{j1}(\xi) & -g_j(\xi) \\ g_j(\xi) & (-1)^j b_{j2}(\xi) \end{pmatrix}, \quad j = 1, 2. \tag{7.15}$$

Let us now evaluate the residue of both sides of (7.14) at $\xi = -\xi^{R(j)}$. This can be done by multiplying them by $\xi + \xi^{R(j)}$ and finding the limits when $\xi \to -\xi^{R(j)}$. By definition of $p(\xi)$, the left-hand side of the resulting equation is zero. The residue of the displacement vector at the Rayleigh pole $\xi = -\xi^{R(j)}$ is

$$\operatorname*{Res}_{\xi=-\xi^{R(j)}} \widehat{\mathbf{u}}^{(j)}(\xi) = \lim_{\xi\to-\xi^{R(j)}}(\xi + \xi^{R(j)})\widehat{\mathbf{u}}^{R(j)}(\xi) = iD^{(j)R}\mathbf{v}^{(j)R}, \quad j = 1, 2. \tag{7.16}$$

Therefore, (7.14) leads to

$$iD^{(j)R}\mathbf{v}^{(j)R} + \frac{1}{R_j'(-\xi^{R(j)})}\widehat{B}^{(j)-}(-\xi^{R(j)})\widehat{\mathbf{t}}(-\xi^{R(j)}) = \mathbf{0}, \quad j = 1, 2, \tag{7.17}$$

where $R_j'(\xi)$ is the derivative of the Rayleigh function $R_j(\xi)$. It follows that, for each medium $I^{(j)}$, $j = 1, 2$, the respective Rayleigh diffraction coefficient can be expressed via $\widehat{\mathbf{t}}(\xi)$ as

$$D^{(j)R} = \frac{i[\widehat{B}_{11}^{(j)-}(-\xi^{R(j)})\widehat{t}_1(-\xi^{R(j)}) + \widehat{B}_{12}^{(j)-}(-\xi^{R(j)})\widehat{t}_2(-\xi^{R(j)})]}{R_j'(-\xi^{R(j)})v_1^{(j)R}}, \quad j = 1, 2. \tag{7.18}$$

8. Numerical results. We have developed a FORTRAN90 program for computing the diffraction coefficient $D^{i(j)}(\theta)$ using (7.6), where the displacements and tractions are evaluated at $\xi = -\kappa_i^{(j)}\cos\theta$, with θ being an observation angle.

As has been discussed above, $\widehat{\mathbf{u}}^{(j)}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ are both singular at the GE pole $\xi = \xi^{inc}$. At the real angles $\theta = \theta^{sh}$, which satisfy the equation

$$\kappa_i^{(j)}\cos(\theta^{sh}) = -\kappa_n^{(m)}\cos(\theta^{inc}), \tag{8.1}$$

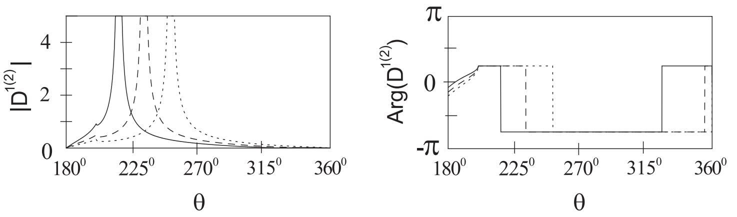


FIG. 8.1. *The longitudinal diffraction coefficient in the lower medium. The solid line corresponds to the angle of incidence of $\theta^{inc} = 30^{o}$, the dashed line to* 50^{o}, *and the dotted line to* 70^{o}.

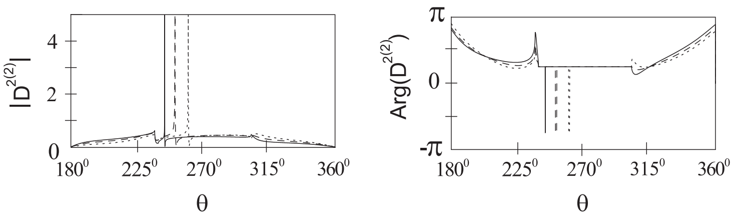


FIG. 8.2. *The transverse diffraction coefficient in the lower medium. The key is as above.*

the estimates of $|D^{i(j)}(\theta)|$ (incorrect near these angles) are infinite and the estimates of $\text{Arg}[D^{i(j)}(\theta^{sh})]$ experience a π jump. When $|\kappa_n^{(m)} \cos(\theta^{inc})/\kappa_i^{(j)}| > 1$, (8.1) has no real valued solutions and $D^{i(j)}(\theta)$ are correct estimates, continuous at all observation angles. The solution of (8.1) is

$$\theta^{sh} = \pi + (-1)^j \arccos\left(\frac{\kappa_n^{(m)}}{\kappa_i^{(j)}} \cos(\theta^{inc})\right). \tag{8.2}$$

When (8.2) defines a real valued angle, it is the shadow boundary of either the reflected or refracted wave in the incident medium, or either the transmitted longitudinal or transverse wave in the other medium.

Other special angles, which can be seen on the graphs in this section, are the so-called critical angles θ^{cr}. They describe the boundaries of the regions that support head waves and correspond to the branch points, that is, satisfy the equation

$$\pm\kappa_k^{(l)} = -\kappa_i^{(j)} \cos\theta^{cr}, \qquad l, k = 1, 2. \tag{8.3}$$

These critical angles do not depend on the angle of incidence. Again, the approximation method used in section 7.1 fails in their vicinity, and they show up on the graphs as small blips.

As an illustration, Figures 8.1–8.4 present diffraction coefficients for a semi-infinite crack, which is sandwiched between aluminum and steel.

The incident plane wave is a longitudinal wave, incoming from the aluminum half plane. The amplitudes of the diffraction coefficients are presented on the left and phases on the right. The model parameters are as follows. In medium 1 (aluminum), density $\rho^{(1)} = 2700$ kg/m^3, longitudinal speed $c_1^{(1)} = 6300$ m/s, and shear speed $c_2^{(1)} = 3100$ m/s. In medium 2 (steel), density $\rho^{(2)} = 7800$ kg/m^3, longitudinal speed $c_1^{(2)} = 5900$ m/s, and shear speed $c_2^{(2)} = 3200$ m/s. We can see the geometrical shadow boundaries described by (8.2), where the amplitude of the formally evaluated

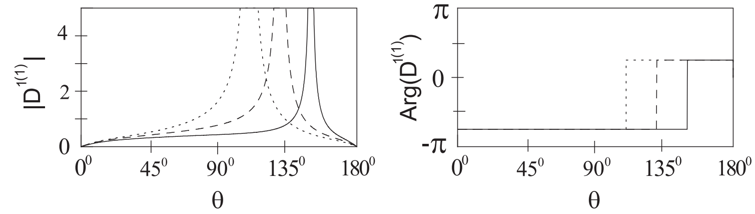


FIG. 8.3. *The longitudinal diffraction coefficient in the upper medium. The key is as in Figure* 8.1.

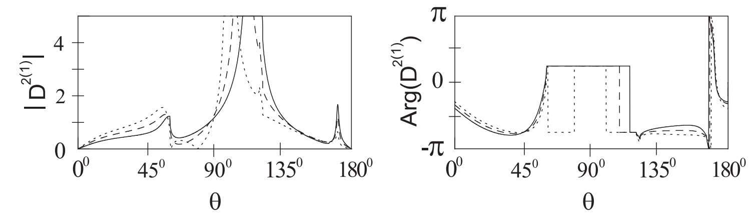


FIG. 8.4. *The transverse diffraction coefficient in the upper medium. The key is as in Figure* 8.1.

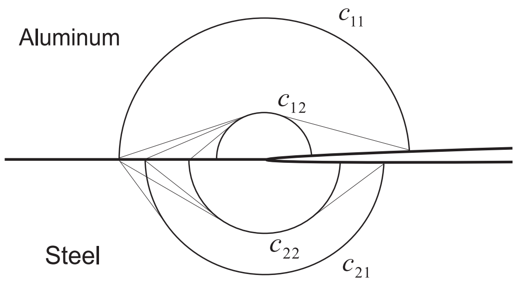


FIG. 8.5. *Head wave fronts.*

diffraction coefficient is infinite and the phase experiences a π jump. Since the critical angles do not depend on the angle of incidence, the corresponding blips appear at the same place for all three curves. The corresponding head wave fronts are shown in Figure 8.5. It can be seen that in the upper medium, the diffracted longitudinal wave is not affected by head waves. This is due to the fact that in aluminum the longitudinal speed is greater than in steel.

In Figure 8.1 the critical angle is $\theta \approx 200^o$. The head wave affects $D^{2(2)}$ at $\theta \approx 237.2^o$, $\theta \approx 239.5^o$, and $\theta \approx 328^o$. In Figure 8.2 only two blips are seen at about $\theta \approx 237^o$ and $\theta \approx 303.6^o$. This is due to the small difference between the first two critical angles $\theta \approx 237.2^o$ and $\theta \approx 239.5^o$. By decreasing both the discretization step and the interval of observation angles, the critical angles separate. The head wave affects $D^{2(1)}$ at $\theta \approx 60.5^o$, $\theta \approx 119.5$, $\theta \approx 121.7^o$, and $\theta \approx 165.6^o$. Again, in Figure 8.4 the critical angles $\theta \approx 119.5^o$ and $\theta \approx 121.7^o$ lie too close to each other. The overall conclusion is that for, say, the 30^o incidence, the longitudinal tip diffracted waves, which propagate in the upper medium, can be best detected at observation angles between 40^o and 130^o. As the angle of incidence increases, the range of advantageous observation angles shortens (see Figure 8.3).

The Rayleigh diffraction coefficients do not depend on the observation angle and are calculated using (7.18). Note that the Rayleigh speeds for aluminum and steel are $c_{Al} = 2894$ m/s and $c_{St} = 2964$ m/s, respectively, and the bimaterial constant is $\nu_0 = 0.0330434$. For the angle of incidence $\theta^{inc} = 30^{\text{o}}$, in aluminum $D^{(1)R} = -0.1208 - 0.206i$, and in steel $D^{(2)R} = 0.1428 - 0.2128i$; for $\theta^{inc} = 50^{\text{o}}$, $D^{(1)R} = -0.1004 - 0.3094i$ and $D^{(2)R} = 0.1285 - 0.3272i$; and for $\theta^{inc} = 70^{\text{o}}$, $D^{(1)R} = -0.0517 - 0.4005i$ and $D^{(2)R} = 0.0784 - 0.4285i$.

We finish this section by discussing the code testing. As already mentioned, the critical angles and positions of the shadow boundaries of reflected and refracted waves, when calculated independently, agree with the above graphs. We also know the phase (either $\pi/4$ or $-3\pi/4$) on those portions of the wave front which are not occupied by the head waves—these are correct. Another stringent internal test is to evaluate the left-hand side of (4.6) on the interval $\xi \epsilon [-1, 1]$. The right-hand side of this equation is known. The left-hand side has been computed using our code. The maximum relative error of each component has been found to be 2%, which is satisfactory. Finally, one could make another check by considering the limiting case of the two identical half planes: In this case, the off-diagonal terms in (4.4) become zero, and (4.3) reduces to the equivalent of the decoupled Wiener–Hopf equations for the crack opening displacements as found in the studies of the isotropic case, e.g., [11]. However, no numerical check of this nature has been conducted, because, when both half planes are the same, (4.7) contains a dipole instead of a pole, and although one could use this representation, it would require additional careful programming: In (4.7) one would have to use the sum of the two poles instead of the product. Even then, numerical difficulties would still arise for nearly identical top and bottom media.

9. Conclusions. We have developed a semianalytical approach to calculating diffraction coefficients for a semi-infinite crack sandwiched between two different isotropic media. We have introduced a stable numerical scheme for solving the resulting system of integral equations, (5.7). Our main achievement has been to produce a fast computer code, which is applicable to any pair of (sufficiently different) isotropic materials which do not support the Stoneley wave and are irradiated by a plane wave incident from either medium. The incident wave can be longitudinal or transverse and incoming at an arbitrary angle. The absence of the Stoneley wave does not constitute a serious restriction, since this case is ubiquitous in applications. Nevertheless, we plan to publish another paper modeling materials where the Stoneley wave is present too. As an illustration, we have presented plots of diffraction coefficients for a crack sandwiched between aluminum and steel.

Appendix A. The two dimensional Green's tensor and Green's stress tensor. Since the incident wave can be considered as radiated by a line load, both the two dimensional Green's tensor and Green's stress tensor can be represented in terms of the Hankel function of the first kind of the zeroth order, $H_0 \equiv H_0^{(1)}$, and its derivatives (see [11], [21]), so that at any observation point $\mathbf{x}$ we have

$$-4i\mu^{(j)}u^{G(j)}_{ik}(\mathbf{x}) = \frac{1}{[k_2^{(j)}]^2}\left[-H_0(k_1^{(j)}r) + H_0(k_2^{(j)}r)\right]_{,ik} + H_0(k_2^{(j)}r)\delta_{ik},$$

$$-4i\sigma^{G(j)}_{2ik}(\mathbf{x}) = \left\{1 - 2\left[\frac{c_2^{(j)}}{c_1^{(j)}}\right]^2\right\}\left[H_0(k_1^{(j)}r)\right]_{,k}\delta_{2i} - \frac{2}{(k_2^{(j)})^2}\left[H_0(k_1^{(j)}r) - H_0(k_2^{(j)}r)\right]_{,2ik}$$

$$+\left[H_0(k_2^{(j)}r)\right]_{,2}\delta_{ik} + \left[H_0(k_2^{(j)}r)\right]_{,i}\delta_{2k}, \tag{A.1}$$

where $\mu^{(j)},\ j=1,2$, is the shear modulus in medium $I^{(j)}$; r is the distance to the origin; and an index, say k, after the comma refers to differentiation with respect to the corresponding spatial variable x_k. Applying operations of dilatation and rotation to both sides of (A.1) gives us

$$\begin{aligned}
&4i\mu^{(j)}u_i^{G(j)1}(\mathbf{x}) = \left[\frac{c_2^{(j)}}{c_1^{(j)}}\right]^2 \left[H_0(k_1^{(j)}r)\right]_{,i}\,,\\
&-4i\sigma_i^{G(j)1}(\mathbf{x}) = (k_1^{(j)})^2\left\{\left[2\left(\frac{c_2^{(j)}}{c_1^{(j)}}\right)^2-1\right]H_0(k_1^{(j)}r)\delta_{2i}+\frac{2}{(k_2^{(j)})^2}\left[H_0(k_1^{(j)}r)\right]_{,2i}\right\},\\
&-4i\mu^{(j)}u_i^{G(j)2}(\mathbf{x}) = \left[H_0(k_2^{(j)}r)\right]_{,1}\delta_{2i}-\left[H_0(k_2^{(j)}r)\right]_{,2}\delta_{1i},\\
&-4i\sigma_i^{G(j)2}(\mathbf{x}) = \left[H_0(k_2^{(j)}r)\right]_{,12}\delta_{2i}-\left[H_0(k_2^{(j)}r)\right]_{,22}\delta_{1i}+\left[H_0(k_2^{(j)}r)\right]_{,1i}\,.
\end{aligned}$$
(A.2)

Appendix B. The extinction theorem. The extinction theorems are easily proved for finite sources and obstacles using Green's theorem. Difficulties arise when the incident waves are plane and obstacles infinite. One approach to dealing with this complication is to develop methods such as those offered in [22] and references therein (also see [23]). Below we offer an alternative justification.

Let us focus on the scattered field in the upper plane. Any identity involving the incident field can be established by direct integration. For simplicity of presentation, we omit the superscript $^{(1)}$. Then solving the Fourier transform of the equations of motion for the elastic solid gives

$$\widehat{\mathbf{u}}^{sc}(\xi,x_2)=A(\xi)(-\xi,\gamma_1)^Te^{ik_1\gamma_1x_2}+B(\xi)(\gamma_2,\xi)^T\mathrm{e}^{ik_1\gamma_2x_2},\quad x_2>0, \tag{B.1}$$

where $A(\xi)$ and $B(\xi)$ are unknown. The solutions proportional to $\exp[-ik_1\gamma_ix_2]$, $i=1,2$, are rejected because they do not satisfy the radiation conditions: Either they are incoming from infinity or else, when we move the branches off the real axis (see Figure C.1 below) as $x_2\to\infty$, they become unbounded. It follows that on the top face of the crack, $x_2=0^+$, we have

$$\widehat{\mathbf{u}}^{sc}(\xi,0^+)=A(\xi)(-\xi,\gamma_1)^T+B(\xi)(\gamma_2,\xi)^T. \tag{B.2}$$

It can easily be verified that a similar formula holds for the traction related vector $\widehat{\mathbf{t}}^{sc}$ (see (4.1)),

$$\widehat{\mathbf{t}}^{sc}(\xi,0^+)=-i[A(\xi)(2\xi\gamma_1,2\xi^2-\kappa_2^2)^T+B(\xi)(2\xi^2-\kappa_2^2,-2\xi\gamma_2)^T]. \tag{B.3}$$

The solution to the Fourier transform of the equations of motion, which is valid in both half planes, is

(B.4)
$$\widehat{\mathbf{u}}^{sc}(\xi,x_2)=A_1(\xi)(-\xi,\gamma_1\mathrm{sgn}(x_2))^Te^{ik_1\gamma_1|x_2|}+B_1(\xi)(\gamma_2\mathrm{sgn}(x_2),\xi)^T\mathrm{e}^{ik_1\gamma_2|x_2|},\ x_2>0,$$

where we have

$$\begin{aligned}
2\kappa_2^2\gamma_1A_1(\xi)&=-2\xi\gamma_1\mathrm{sgn}(x_2)\widehat{u}_1^{sc}(\xi,0^+)+(\kappa_2^2-2\xi^2)\widehat{u}_2^{sc}(\xi,0^+)\\
&\quad+i[\xi\widehat{t}_1^{sc}(\xi,0^+)-\gamma_1\mathrm{sgn}(x_2)\widehat{t}_2^{sc}(\xi,0^+)],\\
2\kappa_2^2\gamma_2B_1(\xi)&=(\kappa_2^2-2\xi^2)\widehat{u}_1^{sc}(\xi,0^+)+2\xi\gamma_2\mathrm{sgn}(x_2)\widehat{u}_2^{sc}(\xi,0^+)\\
&\quad-i[\gamma_2\mathrm{sgn}(x_2)\widehat{t}_1^{sc}(\xi,0^+)+\xi\widehat{t}_2^{sc}(\xi,0^+)].
\end{aligned} \tag{B.5}$$

Substituting (B.2) and (B.3) into (B.5) yields

$$(A_1(\xi), B_1(\xi)) = (A(\xi), B(\xi))H(x_2). \tag{B.6}$$

Thus, the scattered field defined by (B.4) agrees with (B.1) in the upper half plane and is identically zero in the fictitious half plane $x_2 < 0$. The inverse transform of (B.4) leads to the extinction theorem (2.1) for the scattered field.

Appendix C. Singularities in (3.11). Let us describe all singularities of functions that appear in (3.11). First, in view of (3.10), the left-hand side of (3.11) has simple poles at $\xi = \xi^{inc} + i0$ and $\xi = \xi^{inc} - i0$, which correspond to the incident and reflected bulk waves, respectively. They give rise to GE bulk waves.

Second, it is easy to check that the determinant of the matrix $\widehat{A}^+(\xi)$, $|\widehat{A}^+(\xi)|$ is a product of two Rayleigh functions, $R_1(\xi)$ and $R_2(\xi)$,

$$R_j(\xi) = a_j^2(\xi) + 4\xi^2\gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi), \tag{C.1}$$

where the subscript $j = 1, 2$ refers to medium $I^{(j)}$ (see, e.g., [21]). Thus, the solutions $\xi = \pm\xi^{R(j)}$ of the equation $|\widehat{A}^+(\xi)| = 0$ are zeros of $R_1(\xi)$ and $R_2(\xi)$ and can be shown to be simple (distinct). The zeros $\xi = -\xi^{R(j)}$ are known to give rise to the outgoing Rayleigh surface waves.

It is equally easy to check that $|\widehat{A}^-(\xi)|$ is the well-known Stoneley function

$$S(\xi) = \mu^2 R_1(\xi)h_2(\xi) + R_2(\xi)h_1(\xi) + \mu[b_{11}(\xi)b_{22}(\xi) + b_{21}(\xi)b_{12}(\xi) - 2g_1(\xi)g_2(\xi)] \tag{C.2}$$

(see, e.g., [24]), where we use the notation

$$\begin{aligned} &b_{j1}(\xi) = [\kappa_2^{(j)}]^2\gamma_2^{(j)}(\xi), \qquad b_{j2}(\xi) = [\kappa_2^{(j)}]^2\gamma_1^{(j)}(\xi),\\ &g_j(\xi) = \xi[2\gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi) - a_j(\xi)], \qquad h_j(\xi) = \gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi) + \xi^2, \quad j = 1, 2. \end{aligned} \tag{C.3}$$

In general, the zero of $S(\xi) = 0$ (which is also simple) can give rise to an outgoing Stoneley wave. Using Cagniard's method (see [24]) we have established that for the set of parameters used in this paper such a solution does not exist, and therefore no Stoneley surface wave runs between materials under study. We remark in passing that this situation is common, and Cagniard [24] refers to the Stoneley wave as "a rather special phenomenon," meaning that it exists only in narrow ranges of material parameters.

To continue, both matrices $\widehat{A}^{\pm}(\xi)$ involve multivalued radicals $\gamma_i^{(j)}(\xi)$ defined below (3.14). In order to render the matrices single valued we introduce the branch cuts $C_i^{(j)\mp}$, $i, j = 1, 2$, which run between branch points $\pm\kappa_i^{(j)}$, defined below (3.9), and $\pm\infty$, respectively. Let us apply the limiting absorption principle and replace $\kappa_i^{(j)}$ by $\kappa_i^{(j)} + i\epsilon_1$, $\epsilon_1 > 0$. This shifts the branch cuts away from the real axis as indicated in Figure C.1, and when performing the inverse Fourier transform, the corresponding singularities give rise to the waves, which satisfy the radiation condition, that is, are outgoing to infinity.

The radicals $\gamma_i^{(j)}(\xi)$ can be factorized so that we have

$$\gamma_i^{(j)}(\xi) = \gamma_i^{(j)+}(\xi)\gamma_i^{(j)-}(\xi), \tag{C.4}$$

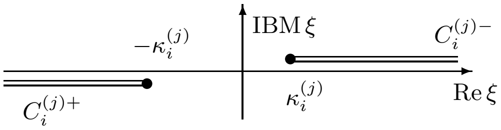


Fig. C.1. *The branch cuts of* $\gamma_i^{(j)}(\xi)$.

where we use the notation

$$\gamma_i^{(j)+}(\xi) = \sqrt{\kappa_i^{(j)} + \xi}, \quad \gamma_i^{(j)-}(\xi) = \gamma_i^{(j)+}(-\xi), \tag{C.5}$$

and only the cut $C_i^{(j)+}$ is required to render $\gamma_i^{(j)+}(\xi)$ single-valued. The radical $\gamma_i^{(j)-}(\xi)$ is rendered single-valued by $C_i^{(j)-}$. Note that if $\xi = -\zeta$ lies on the branch cut $C_i^{(j)+}$, the definition implies that ξ lies on the upper side of the cut, so that $\xi = -\zeta + i0$. It follows that $\gamma_i^{(j)+}(-\zeta)$ is well defined. As ξ tends to $-\zeta$ from below the branch cut, we have

$$\gamma_i^{(j)+}(\xi) \to -\gamma_i^{(j)+}(-\zeta) \equiv \gamma_i^{(j)+}(-\zeta - i0). \tag{C.6}$$

Note that in the main text we drop a combination of subscript and superscript ${}_1^{(1)}$ in the symbol for the longest branch cut $C_1^{(1)\pm}$. To summarize, the known functions in (3.11) involve two GE poles $\xi^{inc} \pm i0$, two Rayleigh poles $-\xi^{R(1)}$ and $-\xi^{R(2)}$, as well as two branch cuts $C^{\pm}$.

Appendix D. Auxiliary functions and vectors. Let us now describe auxiliary functions used in the main text. Let four branch points be sorted in order of the descending moduli, with $\kappa_1 = \min\{\kappa_i^{(j)}\}$ and $\kappa_4 = \max\{\kappa_i^{(j)}\}$, $i, j = 1, 2$, denoting the corresponding radicals $\gamma_i^{\pm}(\xi) = \sqrt{\kappa_i \pm \xi}$ and the respective branch cuts $C_i^+ = \{\xi : \xi \le -\kappa_i\}$ and $C_i^- = \{\xi : \xi \ge \kappa_i\}$, $i = 1, 2, 3, 4$.

For each pair of numbers $a \neq -\kappa_i$ and ξ, let us introduce the auxiliary functions $H_i^{\pm}(\xi, a)$,

$$H_i^+(\xi, a) = \frac{\gamma_i^+(a) - \gamma_i^+(\xi)}{2\gamma_i^+(a)} \left[\frac{a_0 - \gamma_i^+(a)}{a_0 + \gamma_i^+(\xi)}\right]^{2n+1}, \qquad H_i^-(\xi, a) = H_i^+(-\xi, -a), \tag{D.1}$$

where $n \ge 3$ and a_0 lies far away from the branch cut C_i^+. To be specific, let $a_0 = 1+i$. Then the function H_i^+ has the following properties:

- It has no poles in ξ.
- The branch cut C^+ renders it single-valued.
- For any $\xi \in C_i^+$ we have

$$H_i^+(\xi + i0, \xi) = 0 \quad \text{and} \quad H_i^+(\xi - i0, \xi) = 1, \tag{D.2}$$

  where, as the above notation suggests, $H_i^+(\xi + i0, a)$ and $H_i^+(\xi - i0, a)$ are values of $H_i^+(\xi, a)$ evaluated on the upper and lower sides of C_i^+, respectively. (We recall that, unless stated otherwise, ξ lies on the positive side of the cut.)
- For any $\xi \in C_i^-$ such that Re $\zeta \gg 1$ we have

$$H_i^+(-\xi + i0, a) - H_i^+(-\xi - i0, a) \sim \frac{\text{constant}}{[\gamma_i^+(-\xi)]^{2n}}. \tag{D.3}$$

Let us now define branch cuts

$$\begin{aligned}
&C^-_{ii+1} = \{\xi : \kappa_i \le \xi \le \kappa_{i+1}\}, \qquad i = 1, 2, 3,\\
&C^-_{45} = \{\xi : \xi \ge \kappa_4\},\\
&C^+_{ii+1} = \{\xi : -\kappa_{i+1} \le \xi \le -\kappa_i\}, \qquad i = 1, 2, 3,\\
&C^+_{45} = \{\xi : \xi \le -\kappa_4\},
\end{aligned} \tag{D.4}$$

and introduce the auxiliary functions

$$\begin{aligned}
&F^\pm_i(\xi, a) = H^\pm_i(\xi, a) - H^\pm_{i+1}(\xi, a), \quad i = 1, 2, 3,\\
&F^\pm_4(\xi, a) = H^\pm_4(\xi, a).
\end{aligned} \tag{D.5}$$

It is easy to see that each function $F^\pm_i(\xi, a)$ has a branch cut $C^\pm_{ii+1}$, $i = 1, 2, 3, 4$, and $F^\pm_i(\xi, \xi) = 1$ on the negative side of its branch cut and 0 everywhere else. Let us introduce a function $G(\xi)$ as

$$G(\xi) = \left(\sqrt{1+\xi} + \sqrt{\kappa^{(1)}_2 + \xi}\right)^2, \tag{D.6}$$

which is real outside the interval $(-\kappa^{(1)}_2, -1)$ and $O(\xi)$ at infinity.

Appendix E. Auxiliary relationships. Let us determine eight scalar constants $D^\pm_i$, $i = 1, 2, 3, 4$, introduced in (6.1). Let us show that they are linearly dependent, and therefore that the total number of unknowns can be decreased by four. Let us do this by analyzing the asymptotic behavior of both sides of (6.2). As $\xi \to \infty$, matrix $M(\xi) \to M_\infty$,

$$M_\infty = \begin{pmatrix} 0_2 & m^-_\infty I_2 \\ m^+_\infty I_2 & 0_2 \end{pmatrix}, \tag{E.1}$$

where matrices 0_2 and I_2 denote the zero and identity 2×2 matrices, respectively; $m^\pm_\infty$ are known constants,

$$\begin{aligned}
&m^-_\infty = -\frac{1}{8}\left\{\frac{[\kappa^{(1)}_2]^2}{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2} + \mu\frac{[\kappa^{(2)}_2]^2}{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2}\right\},\\
&m^+_\infty = -8[S_\infty]^{-1}\left([\kappa^{(1)}_2]^2\{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2\} + \mu[\kappa^{(2)}_2]^2\{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2\}\right);
\end{aligned} \tag{E.2}$$

and we use the notation $S_\infty = \lim_{\xi\to\infty} S(\xi)/\xi^2$. Using the Stoneley function $S(\xi)$ defined in (C.2), we find

$$\begin{aligned}
S_\infty = {}&\mu^2\{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2\}\{[\kappa^{(2)}_2]^2 + [\kappa^{(2)}_1]^2\} + \{[\kappa^{(1)}_2]^2 + [\kappa^{(1)}_1]^2\}\{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2\}\\
&+ 2\mu\{[\kappa^{(1)}_2]^2[\kappa^{(2)}_2]^2 + [\kappa^{(1)}_1]^2[\kappa^{(2)}_1]^2\}.
\end{aligned} \tag{E.3}$$

The right-hand side of (6.2) decays faster than the left-hand side. It can be shown that it has the asymptotic behavior

$$\Delta\mathbf{V}^{inc}(\xi) \to \frac{1}{\xi}, \qquad \xi \to \infty. \tag{E.4}$$

The asymptotic solution of (6.2) can be rewritten using (6.1) as

$$\Delta\mathbf{w}(\xi) = \xi^{-\nu}\mathbf{W}, \quad \nu = \frac{1}{2} \pm i\nu_0. \tag{E.5}$$

Then the Cauchy integral of the latter can be found in [20] to behave as

$$\frac{1}{\pi}\int_1^\infty \frac{\mathbf{W}d\zeta}{\zeta^\nu(\xi+\zeta)} \to \frac{\beta}{\xi^\nu}\mathbf{W}, \quad \xi \to \infty, \tag{E.6}$$

where $\beta = 1/\sin\nu\pi$. Therefore, using (E.1) and (E.4), as $\xi \to \infty$, the system (6.2) becomes

$$\frac{1}{\xi^\nu}(I_4 - \beta M_\infty)\mathbf{W} = \mathbf{0}. \tag{E.7}$$

The matrix in the above equation must have a zero determinant,

$$\det(I_4 - \beta M_\infty) = \begin{vmatrix} I_2 & -\beta m_\infty^- I_2 \\ -\beta m_\infty^+ I_2 & I_2 \end{vmatrix} = 1 - \beta^2 m_\infty^+ m_\infty^-. \tag{E.8}$$

This determines β, and hence, by its definition, the parameter ν_0:

$$\beta = \frac{1}{\sin\left(\frac{1}{2} \pm i\nu_0\right)\pi} = \frac{1}{\cosh(\pi\nu_0)} = \frac{1}{\sqrt{m_\infty^+ m_\infty^-}}. \tag{E.9}$$

Using one of the Dundurs parameters, which can be represented as

$$\beta_D = \frac{1 - [\kappa^{(2)}]^2 + \mu([\kappa^{(1)}]^2 + 1)}{\mu[\kappa^{(2)}]^2([\kappa^{(1)}]^2 - 1) + [\kappa^{(1)}]^2([\kappa^{(2)}]^2 - 1)} \tag{E.10}$$

(see [25]), as well as expressions (E.2) for $m_\infty^\pm$, we can write

$$\tanh(\pi\nu_0) = \beta_D. \tag{E.11}$$

Now it can be shown that there are two linearly independent vectors $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ such that $(I_4 - \beta M_\infty)\mathbf{W} = \mathbf{0}$. They are

$$\mathbf{W}^{(1)} = \begin{pmatrix} m \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{W}^{(2)} = \begin{pmatrix} 0 \\ m \\ 0 \\ 1 \end{pmatrix}, \quad m = \sqrt{\frac{m_\infty^-}{m_\infty^+}}. \tag{E.12}$$

It follows that as $\xi \to \infty$, the vector $\Delta\mathbf{w}$ behaves as

$$\Delta\mathbf{w}(\xi) \to \xi^{-1/2}\left[\left(\frac{A^+}{\xi^{i\nu_0}} + \frac{A^-}{\xi^{-i\nu_0}}\right)\mathbf{W}^{(1)} + \left(\frac{B^+}{\xi^{i\nu_0}} + \frac{B^-}{\xi^{-i\nu_0}}\right)\mathbf{W}^{(2)}\right], \qquad \xi \to \infty. \tag{E.13}$$

Rewriting the above equation in the form of (6.1) leads us to the following relationship:

$$\begin{aligned} \mathbf{D}^+ &= (A^+ + A^-)\mathbf{W}^{(1)} + (B^+ + B^-)\mathbf{W}^{(2)}, \\ \mathbf{D}^- &= -i[(A^+ - A^-)\mathbf{W}^{(1)} + (B^+ - B^-)\mathbf{W}^{(2)}]. \end{aligned} \tag{E.14}$$

Substituting (E.12) into (E.14) gives us a simple relationship between the components of $\mathbf{D}^\pm$,

$$D_1^\pm = \sqrt{\frac{m_\infty^-}{m_\infty^+}} D_3^\pm, \qquad D_2^\pm = \sqrt{\frac{m_\infty^-}{m_\infty^+}} D_4^\pm. \tag{E.15}$$

Acknowledgments. We are indebted to Profs. V. A. Borovikov and G. Mishuris for numerous enlightening discussions, and to an anonymous referee for insightful comments on the extinction theorems.

REFERENCES

[1] P. Bredif, C. Poideving, and O. Dupond, *A phased array technique for crack characterization*, in Proceedings of ECNDT 2006 (the European Conference on Nondestructive Testing); available online from http://www.ndt.net/article/ecndt2006/doc/Th.1.1.2.pdf.

[2] R. K. Chapman, J. Pearce, S. Burch, L. Fradkin, and M. Toft, *Recent in-house developments in theoretical modelling of ultrasonic inspection*, Insight, 49 (2007), pp. 93–97.

[3] L. A. Ahlberg and B. R. Tittmann, *Measurement techniques in elastic wave scattering experiments*, in Ultrasonics Symposium Proceedings, IEEE, New York, 1980, Vol. 2, pp. 842–846.

[4] B. R. Tittmann, *Scattering of elastic waves from simple defects in solids, A review*, Wave Motion, 5 (1983), pp. 299–306.

[5] K. M. Jaleel, N. N. Kishore, and V. Sundararajan, *Finite-element simulation of elastic-wave propagation in orthotropic composite-materials*, Materials Evaluation, 51 (1993), pp. 830–838.

[6] P. A. Lewis, J. A. G. Temple, E. J. Walker, and G. R. Wickham, *Calculation for diffraction coefficients for a semi-infinite crack embedded in an infinite anisotropic linearly elastic body*, Proc. R. Soc. Lond. A., 454 (1998), pp. 1781–1803.

[7] A. K. Gautesen, *Scattering by elastic quarter space*, Wave Motion, 7 (1985), pp. 557–568.

[8] A. K. Gautesen, *A geometrical theory of diffraction for crack-opening displacements*, Wave Motion, 10 (1988), pp. 393–404.

[9] A. K. Gautesen, *Scattering of a Rayleigh wave by an elastic wedge whose angle is greater than* 180 *degrees*, ASME J. Appl. Mech., 61 (2001), pp. 476–479.

[10] A. K. Gautesen, *On scattering of an SH-wave by a corner comprised of two different elastic materials*, Mech. Mat., 35 (2003), pp. 407–414.

[11] J. D. Achenbach, A. K. Gautesen, and H. McMaken, *Ray Methods for Waves in Elastic Solids: With Applications to Scattering by Cracks*, Pitman, New York, 1982.

[12] M. Lax, *Multiple scattering of waves.* II. *The effective field in dense systems,* Phys. Rev., 85 (1952), pp. 621–629.

[13] V. Kubzina, *Ultrasonic Phenomena with Application to Nondestructive Evaluation*, Ph.D. thesis, Department of Electrical, Electronic and Communications Engineering, London South Bank University, London, 2008.

[14] M. Comninou, *An overview of interface cracks*, Eng. Fracture Mech., 37 (1990), pp. 197–208.

[15] M. L. Williams, *The stresses around a fault or crack in dissimilar media*, Bull. Seismol. Soc. Amer., 49 (1959), pp. 199–404.

[16] F. Erdogan, *Stress distribution in a nonhomogeneous elastic plane with cracks*, J. Appl. Mech., 30 (1963), pp. 232–237.

[17] F. Erdogan, *Stress distribution in bonded dissimilar materials with cracks*, J. Appl. Mech., 32 (1965), pp. 403–410.

[18] J. R. Rice and G. C. Sih, *Plane problems of cracks in dissimilar media*, J. Appl. Mech., 32 (1965), pp. 418–423.

[19] Y. A. Antipov, *An exact solution of the* 3*D-problem of an interface semi-infinite plane crack*, J. Mech. Phys. Solids, 47 (1999), pp. 1051–1093.

[20] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[21] J. D. Achenbach, *Wave Propagation in Elastic Solids*, North–Holland, Amsterdam, 1973.

[22] A. Charalambopoulos, D. Gintides, and K. Kiriaki, *Radiation conditions for rough surfaces in linear elasticity*, Quart. J. Mech. Appl. Math., 55 (2002), pp. 421–441.

[23] J. A. DeSanto, *Exact boundary integral equations for scattering of scalar waves from perfectly reflecting infinite rough surfaces*, Wave Motion, 45 (2008), pp. 918–926.

[24] L. Cagniard, *Reflection and Refraction of Progressive Seismic Waves*, translated and revised by E. A. Flinn and C. H. Dix, McGraw–Hill, New York, 1962.

[25] J. Dundurs, *Effect of elastic constants on stress in a composite under plane deformation,* J. Compos. Mater., 1 (1967), pp. 310–322.

SIAM J. Appl. Math.
Vol. 69, No. 5, pp. 1309–1333

DIFFRACTION BY A SEMI-INFINITE INTERFACIAL CRACK SANDWICHED BETWEEN TWO ISOTROPIC HALF PLANES*

V. KUBZINA[†‡], A. K. GAUTESEN[§], AND L. JU. FRADKIN[†]

Abstract. This paper addresses the canonical two dimensional problem of diffraction of the plane wave by a semi-infinite interfacial crack sandwiched between two isotropic solids. We restrict ourselves to a ubiquitous case of solids whose contact boundary does not support the Stoneley wave. Its solution can be used in applications to model diffraction from curved cracks with curvature that is small compared to a wavelength.

Key words. diffraction, interfacial crack, semianalytical approach

AMS subject classifications. 45E10, 78A45, 74J20

DOI. 10.1137/070711517

1. Introduction.

Finding a semianalytical solution to the problem of diffraction by a semi-infinite crack sandwiched between two different solids is a well-known problem in the mathematical theory of diffraction. In this paper we consider a two dimensional case involving two different isotropic half planes. Apart from being mathematically challenging, diffraction problems of this kind are of interest in ultrasonic NDE (nondestructive evaluation), particularly because interfacial cracks are often found in laminated composites. It is well known that ultrasonic inspection of such cracks is a challenging engineering problem, and detection of crack tip diffraction is particularly difficult. As a consequence, the defect size can be underestimated. Nevertheless, the advanced phased array transducers offer an improved performance [1], and models of the underlying diffraction process would allow the NDE inspectors to establish whether, in a given configuration, the amplitude of the edge diffracted echoes could exceed the detection threshold [2].

Over the years purely numerical approaches to this kind of problem based on finite differences, finite elements, or boundary integral techniques proved unreliable, because it is difficult to take into account the singularity condition at the crack tip and thus render a solution unique. It is also difficult to keep adjusting numerical schemes to account for different types of wave interaction [3, 4, 5]. Another well-known line of attack is to reformulate the problem in terms of a system of functional equations and to solve those using a numerical Wiener–Hopf factorization technique (see, e.g., [6]). So far, this approach has also met with numerous numerical difficulties and has produced no entirely satisfactory scheme.

[*] Received by the editors December 20, 2007; accepted for publication (in revised form) October 30, 2008; published electronically February 19, 2009. A related short report is A. K. Gautesen, V. Kubzina, and L. Ju. Fradkin, *Modeling scatter from 2D interfacial crack*, in Review of Progress in Quantitative Nondestructive Evaluation 24A, D. O. Thompson and D. E. Chimenti, eds., Plenum Press, New York, 2005, pp. 35–42.

http://www.siam.org/journals/siap/69-5/71151.html

[†] Waves and Fields Research Group, Electrical, Computer and Communications Engineering Department, Faculty of Engineering, Science and The Built Environment, London South Bank University, London SE1 0AA, UK (vkubzina@tynemarch.co.uk, fradkil@lsbu.ac.uk). The first author had partial financial support from London South Bank University.

[‡] Current address: Tynemarch Systems Engineering Ltd., Crossways House, 54–60 South Street, Dorking, Surrey, RH4 2HQ, UK.

[§] Department of Mathematics, Iowa State University, and Ames Laboratory, Ames, IA 50011 (gautesen@scl.ameslab.gov).

In this paper we follow a semianalytical approach of [7, 8, 9, 10], developed to derive a solution of the elastodynamic equations formulated in terms of displacements, which it reduces to a system of regular integral equations for their Fourier transforms. We start with the elastodynamic integral equation for the displacement based on Green's formula and the extinction theorem for each isotropic half plane and then use operations of dilatation (or divergence) and rotation (or curl) to separate transverse and longitudinal motions. Since the incident wave can be considered as radiated by a line load, we represent the two dimensional free space Green's tensors in terms of the Hankel function of the first kind of the zeroth order and its derivatives. Using the boundary conditions, the Fourier transform of the elastodynamic integral equation is reduced to a system of four functional equations in eight "half unknowns." Then, the problem is reformulated in terms of traction and crack opening displacement, both of which can be decomposed into singular and nonsingular parts. The singular parts relate to the well-known geometricoelastodynamic (GE) body waves. The nonsingular parts constitute new unknowns. By using a Hilbert-type integral transform, the functional equations are transformed into four regular integral equations in four unknowns. In turn, these are solved numerically. In the far field, diffraction body wave coefficients are obtained. The method can be generalized to model transversely isotropic media.

2. The problem statement.

We consider a two dimensional semi-infinite crack (see Figure 2.1) sandwiched between two different isotropic media $I^{(j)}$, where superscript $j = 1$ corresponds to the medium occupying the "upper" half plane and $j = 2$ means the "lower" half plane. Let the crack be irradiated by a longitudinal ($n = 1$) or transverse ($n = 2$) plane wave, which is incident from the medium $I^{(m)}$, $m = 1$ or 2, and propagates there with the speed $c_n^{(m)}$, where m and n are both fixed throughout the paper. Further, let us assume without loss of generality that the longitudinal speed $c_1^{(1)}$ in the medium $I^{(1)}$ is greater than the longitudinal speed $c_1^{(2)}$ in the medium $I^{(2)}$. Let us further introduce a Cartesian base $\{\mathbf{e}_1, \mathbf{e}_2\}$, with $\mathbf{e}_1$ running along the crack surface and $\mathbf{e}_2$ perpendicular to $\mathbf{e}_1$ and pointing into the "upper" medium. In this base, every vector can be presented in terms of the corresponding coordinates, so that every position vector $\mathbf{x} = (x_1, x_2)$, every displacement vector $\mathbf{u} = (u_1, u_2)$, etc.

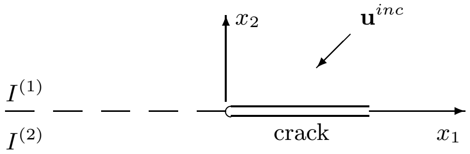


FIG. 2.1. *The problem geometry.*

Let $\mathbf{u}(\mathbf{x}) \exp(-i\omega t)$ be a time harmonic displacement vector in an elastic medium, where t is time, ω is the angular frequency, and the exponential time factor $\exp(-i\omega t)$ is understood but suppressed everywhere below. Using the two dimensional Green's tensors (see Appendix A) and introducing a fictitious bottom medium that has the properties of the upper half space, and a fictitious top medium which has the properties of the bottom half space, the above problem can be recast in the form of a generalized reciprocity relation [11] or the extinction theorem (by analogy with the electromagnetic case—see [12]; the detailed derivation is given in [13]; also see Appendix B). This states that the total displacement for a medium $I^{(j)}$ satisfies the integral equation

$$H[(-1)^{j+1}x_2]u_k^{(j)}(\mathbf{x}) = u_k^{inc(j)}(\mathbf{x}) + (-1)^j \sum_{i=1}^{2} \int_{-\infty}^{-\infty} [\sigma_{2ik}^{G(j)}(x_1 - y_1, x_2)u_i^{(j)}(y_1, 0)$$

$$+ u_{ik}^{G(j)}(x_1 - y_1, x_2)\sigma_{2i}^{(j)}(y_1, 0)]dy_1, \quad j, k = 1, 2, \tag{2.1}$$

where $u^{G(j)}(\mathbf{x} - \mathbf{y})$ and $\sigma^{G(j)}(\mathbf{x} - \mathbf{y})$ are the free space Green's tensor and Green's stress tensor, respectively; $\sigma^{(j)}(\mathbf{x})$ is the stress tensor corresponding to displacement $\mathbf{u}^{(j)}(\mathbf{x})$; $H(x)$ is the Heaviside step function,

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0; \end{cases} \tag{2.2}$$

and in the medium $I^{(j)}$, $j = 1, 2$, the plane wave $\mathbf{u}^{inc(j)}(\mathbf{x})$, which is incident from $I^{(m)}$, is

$$\mathbf{u}^{inc(j)}(\mathbf{x}) = \delta_{mj}\mathbf{d}^{n(m)}e^{-ik_n^{(m)}(p_1^{inc}x_1 - (-1)^m p_2^{inc}x_2)}, \tag{2.3}$$

with δ_{mj}—the Kronecker delta, $\mathbf{p}^{inc} = (p_1^{inc}, p_2^{inc})$—the incoming unit wave vector with $p_2^{inc} > 0$, and $k_n^{(m)} = \omega/c_n^{(m)}$—a wave number (see, e.g., [11]). The longitudinal displacement unit vector $\mathbf{d}^{1(m)}$ is

$$\mathbf{d}^{1(m)} = (-p_1^{inc}, (-1)^m p_2^{inc}), \tag{2.4}$$

and when the motions are transverse the displacement unit vector $\mathbf{d}^{2(m)}$ is

$$\mathbf{d}^{2(m)} = ((-1)^{m+1}p_2^{inc}, -p_1^{inc}). \tag{2.5}$$

To complete the problem statement we require that on the contact boundary, $\{(x_1, x_2) : x_2 = 0,\ x_1 < 0\}$, the displacement and normal stress components be continuous,

$$\begin{aligned} &u_i^{(1)}(x_1, 0) = u_i^{(2)}(x_1, 0), \\ &\sigma_{2i}^{(1)}(x_1, 0) = \sigma_{2i}^{(2)}(x_1, 0), \quad x_1 < 0, \quad i = 1, 2; \end{aligned} \tag{2.6}$$

on the crack $\{(x_1, x_2) : x_2 = 0,\ x_1 > 0\}$ the normal stress components be zero,

$$\sigma_{2i}^{(1)}(x_1, 0) = \sigma_{2i}^{(2)}(x_1, 0) = 0, \quad x_1 > 0, \quad i = 1, 2; \tag{2.7}$$

at infinity, the radiation conditions be satisfied in the form of the limiting absorption principle; and at the crack tip, the mean energy of the diffracted field be bounded. In combination with (2.1), the last condition amounts to requiring that we have

$$\sigma \sim O(r^{-1/2 \pm i\nu_0}), \tag{2.8}$$

with $\nu_0 > 0$ a real bimaterial constant (see, e.g., [14]) and r the distance to the origin $r = \sqrt{x_1^2 + x_2^2}$. The condition (2.8) suggests the oscillatory motions near the crack tip, which is nonphysical. However, this region is often extremely small and in most cases can be ignored. (For further discussion, see Williams [15], Erdogan [16, 17], and Rice and Sih [18]; the related static case has been considered in [19]). Below, if not used as a subscript, $i = \sqrt{-1}$.

3. The functional equations for the Fourier transforms of displacements and stresses in the crack plane. For any field $\varphi(y_1)$ let us define the decomposition

$$\varphi(y_1) = \varphi^+(y_1) + \varphi^-(y_1), \tag{3.1}$$

where the superscripts + and − denote functions that vanish for the negative and positive values of y_1, respectively. Let us use these fields to introduce new "half unknowns," components of the four dimensional vector $\mathbf{v}^{\pm}(y_1)$ given by

$$\begin{aligned}
&v_i^-(y_1) = u_i^{(1)}(y_1, 0) = u_i^{(2)}(y_1, 0), \qquad y_1 < 0,\\
&v_{i+2}^-(y_1) = -\frac{i}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(1)}(y_1, 0) = -\frac{i}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(2)}(y_1, 0), \qquad y_1 < 0,\\
&v_i^+(y_1) = u_i^{(1)}(y_1, 0), \qquad y_1 > 0,\\
&v_{i+2}^+(y_1) = u_i^{(2)}(y_1, 0), \qquad y_1 > 0,\ i = 1, 2.
\end{aligned} \tag{3.2}$$

Let us make use of operators of dilatation (or divergence) and rotation (or curl) and denote the dilatation of any tensor $\phi_{ik}^{(j)}(x_1, x_2)$ by the superscript ^{1} and the rotation by the superscript ^{2}, so that we can write

$$\phi_i^{(j)1}(\mathbf{x}) = [\phi_{i1}^{(j)}(\mathbf{x})]_{,1} + [\phi_{i2}^{(j)}(\mathbf{x})]_{,2}\,, \quad \phi_i^{(j)2}(\mathbf{x}) = [\phi_{i2}^{(j)}(\mathbf{x})]_{,1} - [\phi_{i1}^{(j)}(\mathbf{x})]_{,2}\,. \tag{3.3}$$

Applying the dilatation ($l = 1$) and rotation ($l = 2$) to (2.1) in the half plane where the argument of the Heaviside function is negative ($(-1)^j x_2 > 0$) and using the boundary conditions (2.6) and (2.7), the extinction theorem can be rewritten as

$$\begin{aligned}
&ik_n^{(m)}\delta_{mj}\delta_{nl}e^{-ik_n^{(m)}(x_1p_1^{inc}+(-1)^{m+1}p_2^{inc}x_2)}\\
&\quad + (-1)^j\int_0^\infty [\sigma_1^{G(j)l}(x_1 - y_1, x_2)v_{2j-1}^+(y_1) + \sigma_2^{G(j)l}(x_1 - y_1, x_2)v_{2j}^+(y_1)]dy_1\\
&\quad + (-1)^j\int_{-\infty}^0 \{\sigma_1^{G(j)l}(x_1 - y_1, x_2)v_1^-(y_1) + \sigma_2^{G(j)l}(x_1 - y_1, x_2)v_2^-(y_1)\\
&\qquad\qquad + ik_1^{(1)}\mu^{(1)}[u_1^{G(j)l}(x_1 - y_1, x_2)v_3^-(y_1)\\
&\qquad\qquad + u_2^{G(j)l}(x_1 - y_1, x_2)v_4^-(y_1)]\}dy_1 = 0, \quad j, l = 1, 2,
\end{aligned} \tag{3.4}$$

or in the matrix form as

$$\begin{aligned}
&\int_0^\infty A^+(x_1 - y_1, x_2)\mathbf{v}^+(y_1)dy_1 + \int_{-\infty}^0 A^-(x_1 - y_1, x_2)\mathbf{v}^-(y_1)dy_1\\
&\qquad = -ik_n^{(m)}\mathbf{U}^{inc}e^{ik_n^{(m)}\mathbf{x}\cdot\mathbf{d}^{1(m)}},
\end{aligned} \tag{3.5}$$

where $\mathbf{d}^{1(m)}$ is the displacement unit vector of a longitudinal wave defined in (2.4), $\mathbf{U}^{inc}$ is the four dimensional vector

$$\mathbf{U}^{inc} = [\delta_{1m}\delta_{1n},\ \delta_{1m}\delta_{2n},\ \delta_{2m}\delta_{1n},\ \delta_{2m}\delta_{2n}]^T, \tag{3.6}$$

and 4×4 matrices $A^+(\mathbf{x})$ and $A^-(\mathbf{x})$ are

$$A^+(\mathbf{x}) = \begin{pmatrix} -\sigma_1^{G(1)1}(\mathbf{x}) & -\sigma_2^{G(1)1}(\mathbf{x}) & 0 & 0 \\ -\sigma_1^{G(1)2}(\mathbf{x}) & -\sigma_2^{G(1)2}(\mathbf{x}) & 0 & 0 \\ 0 & 0 & \sigma_1^{G(2)1}(\mathbf{x}) & \sigma_2^{G(2)1}(\mathbf{x}) \\ 0 & 0 & \sigma_1^{G(2)2}(\mathbf{x}) & \sigma_2^{G(2)2}(\mathbf{x}) \end{pmatrix},$$

$$A^-(\mathbf{x}) = \begin{pmatrix} -\sigma_1^{G(1)1}(\mathbf{x}) & -\sigma_2^{G(1)1}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_1^{G(1)1}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_2^{G(1)1}(\mathbf{x}) \\ -\sigma_1^{G(1)2}(\mathbf{x}) & -\sigma_2^{G(1)2}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_1^{G(1)2}(\mathbf{x}) & -ik_1^{(1)}\mu^{(1)}u_2^{G(1)2}(\mathbf{x}) \\ \sigma_1^{G(2)1}(\mathbf{x}) & \sigma_2^{G(2)1}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_1^{G(2)1}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_2^{G(2)1}(\mathbf{x}) \\ \sigma_1^{G(2)2}(\mathbf{x}) & \sigma_2^{G(2)2}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_1^{G(2)2}(\mathbf{x}) & ik_1^{(1)}\mu^{(1)}u_2^{G(2)2}(\mathbf{x}) \end{pmatrix}, \tag{3.7}$$

with dilatations and rotations $u_i^{G(j)l}(\mathbf{x})$ and $\sigma_i^{G(j)l}(\mathbf{x})$ given in Appendix A.

Everywhere below, let the hat $\widehat{\ }$ denote the Fourier transform with respect to the nondimensionalized variable $k_1^{(1)}y_1$, so that for any function $\varphi(y_1)$ we have

$$\widehat{\varphi}(\xi) = k_1^{(1)} \int_{-\infty}^{\infty} \varphi(y_1) e^{ik_1^{(1)}\xi y_1} dy_1. \tag{3.8}$$

Let us then take the Fourier transform of (3.5) and evaluate the result on the boundary. For this purpose, let us multiply (3.5) by $\exp(ik_1^{(1)}\xi x_1)$, integrate it over x_1, and set $x_2 = 0$. Applying the convolution theorem, we obtain

$$\begin{aligned} &\frac{1}{k_1^{(1)}} \left[\int_{-\infty}^{\infty} A^+(\mathbf{x}) e^{ik_1^{(1)}\xi x_1} dx_1\right] \widehat{\mathbf{v}}^+(\xi) \\ &\quad + \frac{1}{k_1^{(1)}} \left[\int_{-\infty}^{\infty} A^-(\mathbf{x}) e^{ik_1^{(1)}\xi x_1} dx_1\right] \widehat{\mathbf{v}}^-(\xi) = p(\xi)\mathbf{U}^{inc}, \end{aligned} \tag{3.9}$$

where, using the notation $\xi^{inc} = \kappa_n^{(m)} p_1$, $\kappa_n^{(m)} = c_1^{(1)}/c_n^{(m)}$, we have

$$p(\xi) = 2\pi\ i\delta(\xi - \xi^{inc}) = \lim_{\epsilon \to 0} \left(-\frac{1}{\xi - \xi^{inc} + i\epsilon} + \frac{1}{\xi - \xi^{inc} - i\epsilon}\right). \tag{3.10}$$

Multiplying the vector equation in (3.9) by $-2i$ and the first and the third scalar equations there by $[\kappa_2^{(1)}]^2$ and $[\kappa_2^{(2)}/\kappa_1^{(2)}]^2$, respectively, we obtain the system of four scalar functional equations

$$\widehat{A}^+(\xi)\widehat{\mathbf{v}}^+(\xi) + \widehat{A}^-(\xi)\widehat{\mathbf{v}}^-(\xi) = p(\xi)\mathbf{v}^{inc}, \tag{3.11}$$

where the vector $\mathbf{v}^{inc}$ is such that

$$\mathbf{v}^{inc} = 2i\left[(\kappa^{(1)})^2\delta_{1m}\delta_{1n},\ \delta_{1m}\delta_{2n},\ (\kappa^{(2)})^2\delta_{2m}\delta_{1n},\ \delta_{2m}\delta_{2n}\right]^T \tag{3.12}$$

with $\kappa^{(j)} = c_1^{(j)}/c_2^{(j)}$; and matrices $\widehat{A}^+(\xi)$ and $\widehat{A}^-(\xi)$ in (3.11) are given by

$$\widehat{A}^+(\xi) = \begin{pmatrix} 2\xi & \frac{a_1(\xi)}{\gamma_1^{(1)}(\xi)} & 0 & 0 \\ -\frac{a_1(\xi)}{\gamma_2^{(1)}(\xi)} & 2\xi & 0 & 0 \\ 0 & 0 & 2\xi & -\frac{a_2(\xi)}{\gamma_1^{(2)}(\xi)} \\ 0 & 0 & \frac{a_2(\xi)}{\gamma_2^{(2)}(\xi)} & 2\xi \end{pmatrix}, \tag{3.13}$$

$$\widehat{A}^-(\xi) = \begin{pmatrix} 2\xi & \frac{a_1(\xi)}{\gamma_1^{(1)}(\xi)} & -\frac{\xi}{\gamma_1^{(1)}(\xi)} & -1 \\ -\frac{a_1(\xi)}{\gamma_2^{(1)}(\xi)} & 2\xi & 1 & -\frac{\xi}{\gamma_2^{(1)}(\xi)} \\ 2\xi & -\frac{a_2(\xi)}{\gamma_1^{(2)}(\xi)} & \frac{\mu\,\xi}{\gamma_1^{(2)}(\xi)} & -\mu \\ \frac{a_2(\xi)}{\gamma_2^{(2)}(\xi)} & 2\xi & \mu & \frac{\mu\,\xi}{\gamma_2^{(2)}(\xi)} \end{pmatrix}, \tag{3.14}$$

with $a_j(\xi) = (\kappa_2^{(j)})^2 - 2\xi^2$, $\gamma_i^{(j)}(\xi) = \sqrt{[\kappa_i^{(j)}]^2 - \xi^2}$, and $\mu = \mu^{(1)}/\mu^{(2)}$.

The dilatation and rotation of the Green's tensor and Green's stress tensor in (3.7) can be expressed in terms of $H_0(k_i^{(j)} r)$, the Hankel function of the first kind of zero order and its derivatives (see Appendix A). Therefore, the matrices $\widehat{A}^\pm(\xi)$ have been evaluated with the help of the following identity:

$$\int_{-\infty}^{\infty} H_0(k_i^{(j)} r) e^{ik_1^{(1)}\xi x_1} dx_1 = \frac{2}{k_1^{(1)}} \frac{1}{\gamma_i^{(j)}} e^{ik_1^{(1)}\gamma_i^{(j)}|x_2|} \tag{3.15}$$

(see, e.g., [20]). Note that (3.11) involves singularities, which are described in Appendix C. Using the definition of plus and minus functions together with (3.8), it is easy to check that $\widehat{\mathbf{v}}^+(\xi)$ and $\widehat{\mathbf{v}}^-(\xi)$ are analytic in the upper and lower halves, respectively, of the complex ξ-plane.

Let us cast (3.11) in another form through multiplying by the matrix $[\widehat{A}^-(\xi)]^{-1}$. Then we obtain the vector functional equation

$$\widehat{\mathbf{v}}^-(\xi) + B^+(\xi)\widehat{\mathbf{v}}^+(\xi) = p(\xi)\mathbf{T}^{inc}, \tag{3.16}$$

where the vector on the right-hand side is

$$\mathbf{T}^{inc} = [\widehat{A}^-(\xi^{inc})]^{-1}\mathbf{v}^{inc}; \tag{3.17}$$

the matrix $B^+(\xi) = [\widehat{A}^-(\xi)]^{-1}\widehat{A}^+(\xi)$ is given by

(3.18)

$$B^+(\xi) = \frac{\mu}{S(\xi)} \cdot$$
$$\begin{pmatrix} b_{21}b_{12} - g_1g_2 + \mu R_1 h_2 & -(b_{21}g_1 + b_{11}g_2) & b_{11}b_{22} - g_1g_2 + \frac{R_2h_1}{\mu} & b_{21}g_1 + b_{11}g_2 \\ b_{22}g_1 + b_{12}g_2 & b_{22}b_{11} - g_1g_2 + \mu R_1 h_2 & -(b_{22}g_1 + b_{12}g_2) & b_{12}b_{21} - g_1g_2 + \frac{R_2h_1}{\mu} \\ -\frac{d_2}{\mu} & \frac{e}{\mu} & \frac{d_2}{\mu} & -\frac{e}{\mu} \\ -\frac{e}{\mu} & -\frac{d_1}{\mu} & \frac{e}{\mu} & \frac{d_1}{\mu} \end{pmatrix},$$

with the Rayleigh functions $R_j(\xi)$, b_{ij}, g_j, $i,j=1,2$, and the Stoneley function $S(\xi)$ defined in Appendix C; and

$$\begin{aligned} d_j(\xi) &= R_2(\xi)b_{1j}(\xi) + \mu R_1(\xi)b_{2j}(\xi), \\ e(\xi) &= R_2(\xi)g_1(\xi) - \mu R_1(\xi)g_2(\xi), \quad j=1,2. \end{aligned} \tag{3.19}$$

In the case under consideration, the equation $|\widehat{A}^-(\xi)|=0$ has no solutions (see Appendix C), and therefore the matrix $B^+(\xi)$ is bounded. Equation (3.16) is a vector functional equation, which has no known analytical solution and is difficult to solve numerically.

4. The functional equations for nonsingular unknowns. Let us reformulate (3.16) as a vector functional equation for nonsingular components of the displacements and stresses in the crack plane. We start by introducing two new unknown vector functions, the crack opening displacement (COD) $\mathbf{u}^{COD}(y_1)$ and $\mathbf{t}(y_1)$, as

$$\begin{aligned} \mathbf{u}^{COD}(y_1) &= \mathbf{u}^{(1)}(y_1,0) - \mathbf{u}^{(2)}(y_1,0), \\ \mathbf{t}(y_1) &= \frac{1}{k_1^{(1)}\mu^{(1)}}\sigma_{2i}^{(1)}(y_1,0). \end{aligned} \tag{4.1}$$

Then using the definition (3.2), their Fourier transforms $\widehat{\mathbf{u}}^{COD}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ may be expressed in terms of components of the old half unknowns $\mathbf{v}^{\pm}(\xi)$ as

$$\begin{aligned} \widehat{u}_l^{COD}(\xi) &= \widehat{v}_l^+(\xi) - \widehat{v}_{2+l}^+(\xi), \\ \widehat{t}_l(\xi) &= i\widehat{v}_{l+2}^-(\xi), \quad l=1,2. \end{aligned} \tag{4.2}$$

The last two equations in (3.16) can be rewritten in terms of the new unknowns $\widehat{\mathbf{u}}^{COD}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ as

$$-i\widehat{\mathbf{t}}(\xi) + \tilde{B}^+(\xi)\widehat{\mathbf{u}}^{COD}(\xi) = p(\xi)\tilde{\mathbf{T}}^{inc}, \tag{4.3}$$

where $p(\xi)$ is defined in (3.10); the matrix $\tilde{B}^+(\xi)$ and vector $\tilde{\mathbf{T}}^{inc}$ are bounded and are

$$\tilde{B}^+(\xi) = \frac{1}{S(\xi)}\begin{pmatrix} -d_2(\xi) & e(\xi) \\ -e(\xi) & -d_1(\xi) \end{pmatrix}, \qquad \tilde{\mathbf{T}}^{inc} = (T_3^{inc}, T_4^{inc})^T. \tag{4.4}$$

The right-hand side of the above equation has two poles, $\xi^{inc}+i0$ and $\xi^{inc}-i0$, which lie infinitely close to each other. Let us isolate one of them by introducing

$$\widehat{\mathbf{s}}(\xi) = -i\widehat{\mathbf{t}}(\xi) - \frac{1}{\xi-\xi^{inc}-i0}\tilde{\mathbf{T}}^{inc}. \tag{4.5}$$

Then the second pole $\xi^{inc}-i0$ can be shifted back to the real axis. For this reason, everywhere below we use ξ^{inc} to mean $\xi^{inc}-i0$. Using (4.5), the functional equation (4.3) can be rewritten as

$$\widehat{\mathbf{s}}(\xi) + \tilde{B}^+(\xi)\widehat{\mathbf{u}}^{COD}(\xi) = -\frac{1}{\xi-\xi^{inc}}\tilde{\mathbf{T}}^{inc}. \tag{4.6}$$

The function $\widehat{\mathbf{u}}^{COD}(\xi)$ is analytic in the upper half plane. Therefore, its domain contains both Rayleigh poles $-\xi^{R(j)}$, $j=1,2$, the incident pole $\xi=\xi^{inc}$, and also

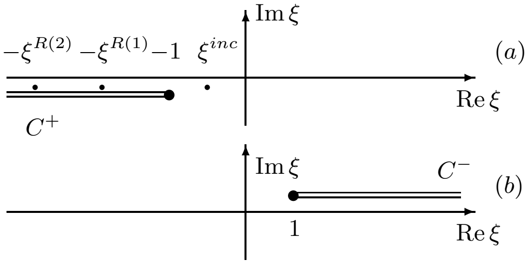


FIG. 4.1. *The poles and branch cuts of* (a) $\widehat{\mathbf{u}}^{COD}(\xi)$ *and* (b) $\widehat{\mathbf{s}}(\xi)$.

the branch cut C^+ (see Figure 4.1(a) and Appendix C). This means that we can decompose $\widehat{\mathbf{u}}^{COD}(\xi)$ as follows:

$$\widehat{\mathbf{u}}^{COD}(\xi) = \frac{G(\xi^{inc})}{G(\xi)}\widehat{\mathbf{w}}^{ge+}(\xi) + \frac{G(\xi)}{[\xi+\xi^{R(1)}][\xi+\xi^{R(2)}]}\widehat{\mathbf{w}}^{+}(\xi), \tag{4.7}$$

where $G(\xi)$ is defined in Appendix D, $\widehat{\mathbf{w}}^{ge+}(\xi)$ has the pole at $\xi = \xi^{inc}$, and $\widehat{\mathbf{w}}^{+}(\xi)$ is a new unknown vector describing the edge diffracted body and surface waves. Note that as $\xi \to \infty$, the function $G(\xi) \sim \xi$, and therefore $\widehat{\mathbf{w}}^{+}(\xi)$ behaves as $\xi\widehat{\mathbf{u}}^{COD}(\xi)$.

The function $\widehat{\mathbf{w}}^{ge+}(\xi)$ can be chosen to be

$$\begin{aligned}\widehat{\mathbf{w}}^{ge+}(\xi) = &-\frac{1}{\xi-\xi^{inc}}\tilde{B}^{-}(\xi^{inc})\tilde{\mathbf{T}}^{inc}\\ &+\frac{1}{\xi-\xi^{inc}}\frac{1}{G(\xi^{inc})}\sum_{k=1}^{4}F_k^{+}(\xi,\xi^{inc})\Delta_k^{+}[G(\xi)\tilde{B}^{-}(\xi)]\big|_{\xi=\xi^{inc}}\tilde{\mathbf{T}}^{inc},\end{aligned} \tag{4.8}$$

where the auxiliary functions $F_k^{+}(\xi,\zeta)$, $k = 1,\dots,4$, are defined in Appendix D; we have

$$\tilde{B}^{-}(\xi) = \frac{1}{R_1(\xi)R_2(\xi)}\begin{pmatrix} -d_1(\xi) & -e(\xi) \\ e(\xi) & -d_2(\xi) \end{pmatrix}, \tag{4.9}$$

and for any function $\Phi(\xi)$ we denote by $\Delta_k^{+}\Phi(\xi)$ the jump of $\Phi(\xi)$ over the branch cut C_k^{+} defined in (D.4) of Appendix D.

Analogously, the function $\widehat{\mathbf{s}}(\xi)$ is analytic in the lower half plane, and we can write the decomposition

$$\widehat{\mathbf{s}}(\xi) = \widehat{\mathbf{w}}^{ge-}(\xi) + \widehat{\mathbf{w}}^{-}(\xi), \tag{4.10}$$

where $\widehat{\mathbf{w}}^{ge-}(\xi)$ is introduced to cancel the extraneous incident pole appearing on the negative side of the branch cut C^- (see Figure 4.1(b)) and $\widehat{\mathbf{w}}^{-}(\xi)$ is a new unknown vector. The function $\widehat{\mathbf{w}}^{ge-}(\xi)$ can be chosen to be

$$\widehat{\mathbf{w}}^{ge-}(\xi) = -\frac{1}{\xi-\xi^{inc}}\sum_{k=1}^{4}F_k^{-}(\xi,\xi^{inc})\Delta_k^{-}[\tilde{B}^{+}(\xi)]\big|_{\xi=\xi^{inc}}\tilde{B}^{-}(\xi^{inc})\tilde{\mathbf{T}}^{inc}, \tag{4.11}$$

where the auxiliary functions $F_k^{-}(\xi,\zeta)$ are defined in Appendix D and $\Delta_k^{-}\Phi(\xi)$ denotes a jump of a function $\Phi(\xi)$ over the branch cut C_k^{-} defined in (D.4). It can be shown that whether or not $\xi = \xi^{inc}$ lies close to the branch cut C^-, we can always use the definition (4.11).

Let us substitute decompositions (4.7) and (4.10) into the functional equation (4.6) to obtain a functional equation for the new unknowns $\widehat{\mathbf{w}}^+(\xi)$ and $\widehat{\mathbf{w}}^-(\xi)$,

$$\widehat{\mathbf{w}}^-(\xi) + M^+(\xi)\widehat{\mathbf{w}}^+(\xi) = \mathbf{V}^{inc-}(\xi), \tag{4.12}$$

where the matrix $M^+(\xi)$ is given by

$$M^+(\xi) = \frac{G(\xi)}{(\xi^{R(1)} + \xi)(\xi^{R(2)} + \xi)}\tilde{B}^+(\xi) \tag{4.13}$$

and the vector $\mathbf{V}^{inc-}(\xi)$ on the right-hand side of (4.12) can be written as

$$\mathbf{V}^{inc-}(\xi) = -\frac{1}{\xi - \xi^{inc}}\tilde{\mathbf{T}}^{inc} - \frac{G(\xi^{inc})}{G(\xi)}\tilde{B}^+(\xi)\widehat{\mathbf{w}}^{ge+}(\xi) - \widehat{\mathbf{w}}^{ge-}(\xi). \tag{4.14}$$

Multiplying (4.12) by the inverse matrix $[M^+(\xi)]^{-1}$, we obtain the functional equation

$$\widehat{\mathbf{w}}^+(\xi) + M^-(\xi)\widehat{\mathbf{w}}^-(\xi) = \mathbf{V}^{inc+}(\xi), \tag{4.15}$$

where $M^-(\xi) = [M^+(\xi)]^{-1}$, so that we have

$$M^-(\xi) = \frac{(\xi^{R(1)} + \xi)(\xi^{R(2)} + \xi)}{G(\xi)R_1(\xi)R_2(\xi)}\begin{pmatrix} -d_1(\xi) & -e(\xi) \\ e(\xi) & -d_2(\xi) \end{pmatrix}, \tag{4.16}$$

and the vector $\mathbf{V}^{inc+}(\xi)$ is

$$\mathbf{V}^{inc+}(\xi) = M^-(\xi)\mathbf{V}^{inc-}(\xi). \tag{4.17}$$

The functional vector equations (4.12) and (4.15) form a system of four functional equations in four unknowns.

5. The system of integral equations. Since the vector $\widehat{\mathbf{w}}^+(\xi)$ introduced in the decomposition (4.7) has no poles and vanishes at infinity, it can be represented as a Hilbert transform

$$\widehat{\mathbf{w}}^+(\xi) = -\frac{1}{2\pi i}\int_1^\infty \frac{\Delta\mathbf{w}^+(\xi')}{\xi' + \xi}d\xi', \tag{5.1}$$

where $\Delta\mathbf{w}^+(\xi)$ is a jump of $\widehat{\mathbf{w}}^+(\xi)$ over the branch cut C^+ (see Figure 4.1(a)); i.e., we have

$$\Delta\mathbf{w}^+(\xi) = \widehat{\mathbf{w}}^+(-\xi + i0) - \widehat{\mathbf{w}}^+(-\xi - i0), \quad \xi > 1. \tag{5.2}$$

Also, the vector $\widehat{\mathbf{w}}^-(\xi)$ introduced in (4.10) can be represented as

$$\widehat{\mathbf{w}}^-(\xi) = -\frac{1}{2\pi i}\int_1^\infty \frac{\Delta\mathbf{w}^-(\xi')}{\xi' - \xi}d\xi', \tag{5.3}$$

where, for $\xi \in C^-$, $\Delta\mathbf{w}^-(\xi)$ is a jump of $\widehat{\mathbf{w}}^-(\xi)$ over the branch cut C^-(see Figure 4.1(b)); i.e., we have

$$\Delta\mathbf{w}^-(\xi) = \widehat{\mathbf{w}}^-(\xi - i0) - \widehat{\mathbf{w}}^-(\xi + i0), \quad \xi > 1. \tag{5.4}$$

Note that $\widehat{\mathbf{w}}^{+}(\xi)$ and $\widehat{\mathbf{w}}^{-}(\xi)$ have the same branches as $\widehat{\mathbf{v}}^{+}(\xi)$ and $\widehat{\mathbf{v}}^{-}(\xi)$, respectively (see, e.g., [9]).

Let us now substitute (5.1) into (4.15), (5.3) into (4.12), and calculate the jumps in the resulting equations across the branch cuts C^+ and C^-, respectively. Then for $\xi > 1$ we have

$$\begin{aligned} &\Delta\mathbf{w}^{+}(\xi) + \Delta M^{-}(-\xi)\widehat{\mathbf{w}}^{-}(-\xi) = \Delta\mathbf{V}^{inc+}(\xi),\\ &\Delta\mathbf{w}^{-}(\xi) + \Delta M^{+}(\xi)\widehat{\mathbf{w}}^{+}(\xi) = \Delta\mathbf{V}^{inc-}(\xi), \end{aligned} \tag{5.5}$$

where Δ denotes the jump over the cut, $\xi \in C^-$, and we use the notation

$$\Delta M^{\pm}(\pm\xi) = 2i\,\mathrm{Im}\,M^{\pm}(\pm\xi), \quad \xi \geq 1. \tag{5.6}$$

In view of (3.10), (4.16), (5.2), (5.4), and the definitions of the functions $F_k^{\pm}(\xi, a)$, $k = 1, \dots, 4$ (see Appendix D), the apparent poles $\xi = \xi^{R(j)}$ and $\xi = \pm\xi^{inc}$ in (5.5) are in fact absent (their residues are zero). Decompositions (4.8) and (4.11) have been chosen in order to achieve this regularization.

Equation (5.5) can be rewritten as the system of coupled integral equations

$$\begin{aligned} &\Delta\mathbf{w}^{+}(\xi) - \frac{1}{\pi}\mathrm{Im}\,[M^{-}(-\xi)]\int_1^\infty \frac{\Delta\mathbf{w}^{-}(\xi')}{\xi+\xi'}d\xi' = \Delta\mathbf{V}^{inc+}(\xi),\\ &\Delta\mathbf{w}^{-}(\xi) - \frac{1}{\pi}\mathrm{Im}\,[M^{+}(\xi)]\int_1^\infty \frac{\Delta\mathbf{w}^{+}(\xi')}{\xi+\xi'}d\xi' = \Delta\mathbf{V}^{inc-}(\xi). \end{aligned} \tag{5.7}$$

6. The numerical scheme. As mentioned above, we seek the solution of (5.7) that allows the tip condition (2.8) to be satisfied, that is, exhibit the asymptotic behavior

$$\Delta\mathbf{w}(\xi) \to \xi^{-1/2}[\mathbf{D}^{+}\cos(\nu_0 \ln\xi) + \mathbf{D}^{-}\sin(\nu_0\ln\xi)], \quad \xi \to \infty, \tag{6.1}$$

where ν_0 is a bimaterial constant; $\mathbf{D}^{\pm}$ are unknown four dimensional constant vectors, and $\Delta\mathbf{w}(\xi) = (\Delta w_1^{+}(\xi), \Delta w_2^{+}(\xi), \Delta w_1^{-}(\xi), \Delta w_2^{-}(\xi))$ is the unknown four dimensional vector function.

Let us rewrite the system of (5.7) as one vector integral equation,

$$\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\int_1^\infty \frac{\Delta\mathbf{w}(\xi')}{\xi'+\xi}d\xi' = \Delta\mathbf{V}^{inc}(\xi), \tag{6.2}$$

where we use the notation

$$\Delta\mathbf{V}^{inc}(\xi) = (\Delta {V_1}^{inc+}(\xi), \Delta {V_2}^{inc+}(\xi), \Delta {V_1}^{inc-}(\xi), \Delta {V_2}^{inc-}(\xi))^T,$$

with $\Delta {V_i}^{inc\pm}(\xi),\ i = 1, 2$, being the right-hand side of (5.5) and $M(\xi)$ being the 4×4 matrix that can be written as

$$M(\xi) = \begin{pmatrix} 0_2 & \mathrm{Im}\,[M^{-}(-\xi)] \\ \mathrm{Im}\,[M^{+}(\xi)] & 0_2 \end{pmatrix}. \tag{6.3}$$

Then there exists a constant L_2 such that for $\xi \geq L_2$ the solution $\Delta\mathbf{w}(\xi)$ exhibits the behavior (6.1) and we can rewrite (6.2) as

$$\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\int_1^{L_2} \frac{\Delta\mathbf{w}(\xi')}{\xi'+\xi}d\xi' - \frac{1}{\pi}M(\xi)[\mathbf{D}^{+}I^{+}(\xi) + \mathbf{D}^{-}I^{-}(\xi)] = \Delta\mathbf{V}^{inc}(\xi), \tag{6.4}$$

where we use the notation

$$
\begin{aligned}
I^+(\xi) &= \frac{1}{2}\left(\int_{L_2}^{\infty} \frac{x^{-1/2+i\nu_0}}{x+\xi}dx + \int_{L_2}^{\infty} \frac{x^{-1/2-i\nu_0}}{x+\xi}dx\right),\\
I^-(\xi) &= \frac{1}{2i}\left(\int_{L_2}^{\infty} \frac{x^{-1/2+i\nu_0}}{x+\xi}dx - \int_{L_2}^{\infty} \frac{x^{-1/2-i\nu_0}}{x+\xi}dx\right).
\end{aligned}
\tag{6.5}
$$

The integral in (6.4) can be approximated using N collocation points $\xi_i,\ i = 1, N$, and weights $W_i,\ i = 1, N$, so that (6.4) may be approximated by

$$
\Delta\mathbf{w}(\xi) - \frac{1}{\pi}M(\xi)\sum_{i=1}^{N}\frac{\Delta\mathbf{w}(\xi_i)}{\xi_i+\xi}W_i - \frac{1}{\pi}M(\xi)[\mathbf{D}^+I^+(\xi) + \mathbf{D}^-I^-(\xi)] = \Delta\mathbf{V}^{inc}(\xi),
\tag{6.6}
$$

which at $\xi = \xi_j,\ j = 1, N$, becomes

$$
\Delta\mathbf{w}(\xi_j) - \frac{1}{\pi}M(\xi_j)\sum_{i=1}^{N}\frac{\Delta\mathbf{w}(\xi_i)}{\xi_i+\xi_j}W_i - \frac{1}{\pi}M(\xi_j)[\mathbf{D}^+I^+(\xi_j) + \mathbf{D}^-I^-(\xi_j)] = \Delta\mathbf{V}^{inc}(\xi_j),
\quad j = 1, \dots, N.
\tag{6.7}
$$

System (6.7) contains $4N$ equations with $4N+8$ unknowns $\Delta w_i(\xi_j),\ i = 1, 4,\ j = 1, N$, and $D_i^{\pm},\ i = 1, 4$, and is thus underdetermined. Four extra equations are provided by the continuity of the solution at the point $\xi_N = L_2$ and can be written as

$$
\Delta\mathbf{w}(\xi_N) = \xi_N^{-1/2}\left[\mathbf{D}^+\cos(\nu_0\ln\xi_N) + \mathbf{D}^-\sin(\nu_0\ln\xi_N)\right].
\tag{6.8}
$$

Four more can be supplied by the relation (E.15) (see Appendix E). The resulting algebraic system of $4N+8$ equations in $4N+8$ unknowns can be written as

$$
M\Delta\mathbf{w} = \mathbf{f},
\tag{6.9}
$$

where the $4N+8$ dimensional vectors are

$$
\begin{aligned}
&\Delta\mathbf{w} = (\Delta w_1(\xi_1), \Delta w_2(\xi_1), \Delta w_3(\xi_1), \Delta w_4(\xi_1), w_1(\xi_2), \dots,\\
&\qquad\qquad \Delta w_4(\xi_N), D_1^+, \dots, D_4^+, D_1^-, \dots, D_4^-),\\
&\mathbf{f} = (\Delta V_1^{inc}(\xi_1), \dots, \Delta V_4^{inc}(\xi_1), \Delta V_1^{inc}(\xi_2), \dots, \Delta V_4^{inc}(\xi_N), 0, 0, 0, 0, 0, 0, 0, 0),
\end{aligned}
\tag{6.10}
$$

and matrix M may be given in a block form as

$$
M =
\begin{pmatrix}
I_4 - \frac{M(\xi_1)W_1}{2\pi\xi_1} & -\frac{M(\xi_1)W_2}{\pi(\xi_1+\xi_2)} & \cdots & -\frac{M(\xi_1)W_{N-1}}{\pi(\xi_1+\xi_{N-1})} & -\frac{M(\xi_1)W_N}{\pi(\xi_1+\xi_N)} & M^D(\xi_1)\\
-\frac{M(\xi_2)W_1}{\pi(\xi_1+\xi_2)} & I_4 - \frac{M(\xi_2)W_2}{2\pi\xi_2} & \cdots & -\frac{M(\xi_2)W_{N-1}}{\pi(\xi_2+\xi_{N-1})} & -\frac{M(\xi_2)W_N}{\pi(\xi_2+\xi_N)} & M^D(\xi_2)\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
-\frac{M(\xi_N)W_1}{\pi(\xi_1+\xi_N)} & -\frac{M(\xi_N)W_2}{\pi(\xi_2+\xi_N)} & \cdots & -\frac{M(\xi_N)W_{N-1}}{\pi(\xi_N+\xi_{N-1})} & I_4 - \frac{M(\xi_N)W_N}{2\pi\xi_N} & M^D(\xi_N)\\
0_4 & 0_4 & \cdots & 0_4 & I_4 & M^{V1}\\
0_4 & 0_4 & \cdots & 0_4 & 0_4 & M^{V2}
\end{pmatrix},
\tag{6.11}
$$

where we use the notation

$$\begin{aligned} &M^D(\xi) = -\frac{1}{\pi} M(\xi) J\left(I^+(\xi), I^-(\xi)\right), \\ &M^{V1} = -J\left(\xi_N^{-1/2}\cos(\nu_0 \ln \xi_N), \xi_N^{-1/2}\sin(\nu_0 \ln \xi_N)\xi_N)\right), \end{aligned} \tag{6.12}$$

with matrix $J\left(f_1(\xi), f_2(\xi)\right)$ given by

$$J\left(f_1(\xi), f_2(\xi)\right) = \begin{pmatrix} f_1(\xi) & f_2(\xi) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_1(\xi) & f_2(\xi) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_1(\xi) & f_2(\xi) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & f_1(\xi) & f_2(\xi) \end{pmatrix}, \tag{6.13}$$

and

$$M^{V2} = \begin{pmatrix} I_2 & -\sqrt{\frac{m_\infty^-}{m_\infty^+}} I_2 & 0_2 & 0_2 \\ 0_2 & 0_2 & I_2 & -\sqrt{\frac{m_\infty^-}{m_\infty^+}} I_2 \end{pmatrix}. \tag{6.14}$$

To evaluate integrals $I^{\pm}(\xi)$ in (6.5), we consider two cases:

- Case 1: $\xi \le L_2$. It is easy to see that we can write

  $$\frac{x}{x+\xi} = \sum_{k=0}^{\infty} (-1)^k \left(\frac{\xi}{x}\right)^k, \tag{6.15}$$

  where the right-hand side is the geometric series with the common ratio $-\xi/x$ such that $|\xi/x| < 1$. Therefore, we have

  $$\int_{L_2}^{\infty} \frac{x^{-1/2 \pm i\nu_0}}{x+\xi} dx = \int_{L_2}^{\infty} x^{-3/2 \pm i\nu_0} \sum_{k=0}^{\infty} (-1)^k \left(\frac{\xi}{x}\right)^k dx. \tag{6.16}$$

  Exchanging in the right-hand side the order of summation and integration and evaluating the resulting integrals, we obtain

  $$\int_{L_2}^{\infty} \frac{x^{-1/2 \pm i\nu_0}}{x+\xi} dx = (L_2)^{-1/2 \pm i\nu_0} \sum_{k=0}^{\infty} \frac{(-1)^k}{0.5 + k \mp i\nu_0} \left(\frac{\xi}{L_2}\right)^k. \tag{6.17}$$

  Substituting (6.17) into (6.5) gives us

  $$\begin{aligned} I^+(\xi) &= \frac{1}{\sqrt{L_2}} \left\{ S_1\left(\frac{\xi}{L_2}\right) \cos(\nu_0 \ln L_2) - \nu_0 S_2\left(\frac{\xi}{L_2}\right) \sin(\nu_0 \ln L_2) \right\}, \\ I^-(\xi) &= \frac{1}{\sqrt{L_2}} \left\{ S_1\left(\frac{\xi}{L_2}\right) \sin(\nu_0 \ln L_2) + \nu_0 S_2\left(\frac{\xi}{L_2}\right) \cos(\nu_0 \ln L_2) \right\}, \end{aligned} \tag{6.18}$$

  where we use the notation

  $$S_1(\xi) = \sum_{k=0}^{\infty} \frac{(-1)^k (0.5+k)}{(0.5+k)^2 + \nu_0^2} \xi^k, \qquad S_2(\xi) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(0.5+k)^2 + \nu_0^2} \xi^k. \tag{6.19}$$

Note that, according to the Leibnitz theorem, the alternating series $S_i(\xi/L_2)$, $i = 1, 2$, converge. The difference between the infinite series and partial sum $S_{iN}(\xi/L_2)$ does not exceed the $(N + 1)$th term. This property can be used to establish the number of terms necessary to achieve the required accuracy. The latter can be improved further if instead of a partial sum S_{iN} we use S_{iN} plus one half of the $(N + 1)$th term. Then the error in $S_i(\xi/L_2)$, $i = 1, 2$, is $O(k^{-i-1})$ rather than $O(k^{-i})$.

- Case 2: $\xi > L_2$. Let us rewrite each term in (6.5) as

$$\int_{L_2}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \int_{0}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx - \int_{0}^{L_2} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx. \tag{6.20}$$

The first integral can be evaluated to give

$$\int_{0}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \frac{\pi}{\cosh \pi\nu_o}\xi^{-1/2\pm i\nu_0} \tag{6.21}$$

(see, e.g., [20]). The integrand of the second term in (6.20) can be represented as

$$\frac{x^{-1/2\pm i\nu_0}}{x+\xi} = \sum_{k=0}^{\infty} \frac{x^{-1/2+k\pm i\nu_0}}{\xi^{k+1}}. \tag{6.22}$$

Integrating both sides of (6.22) over $\xi \in [0, L_2]$ and using (6.21), the integral (6.20) can be represented as

(6.23)
$$\int_{L_2}^{\infty} \frac{x^{-1/2\pm i\nu_0}}{x+\xi}dx = \frac{\pi}{\cosh \pi\nu_o}\xi^{-1/2\pm i\nu_0} - \frac{(L_2)^{0.5\pm i\nu_0}}{\xi}\sum_{k=0}^{\infty} \frac{(-1)^k}{0.5+k\pm i\nu_0}\left(\frac{L_2}{\xi}\right)^k.$$

Therefore, (6.5) becomes

$$\begin{aligned} I^+(\xi) &= -\frac{\sqrt{L_2}}{\xi}\left\{S_1\left(\frac{L_2}{\xi}\right)\cos(\nu_0 \ln L_2) + \nu_0 S_2\left(\frac{L_2}{\xi}\right)\sin(\nu_0 \ln L_2)\right\} \\ &\quad + \frac{\pi}{\sqrt{\xi}\cosh \pi\nu_0}\cos(\nu_0 \ln \xi), \\ I^-(\xi) &= -\frac{\sqrt{L_2}}{\xi}\left\{S_1\left(\frac{L_2}{\xi}\right)\sin(\nu_0 \ln L_2) - \nu_0 S_2\left(\frac{L_2}{\xi}\right)\cos(\nu_0 \ln L_2)\right\} \\ &\quad + \frac{\pi}{\sqrt{\xi}\cosh \pi\nu_0}\sin(\nu_0 \ln \xi). \end{aligned} \tag{6.24}$$

Note that, similarly to the previous case, according to the Leibnitz theorem, the alternating series $S_i(L_2/\xi)$, $i = 1, 2$, converge, and instead of the partial sum S_{iN}, we can utilize S_{iN} plus one half of the $(N + 1)$th term.

The resulting linear system (6.9) can be solved numerically using the LU decomposition subroutines from the LAPACK library. After the unknowns are evaluated at the nodes, system (6.6) can be used to extrapolate the solution for any $\xi \geq 1$. Note that in system (6.9) the integrals $I^{\pm}(\xi)$ are calculated at $\xi \leq L_2$, and therefore when solving this system we utilize only (6.18). When extrapolating, (6.24) is used instead. It is easy to see that, whatever the case, the integrals $I^{\pm}(\xi)$ are real valued.

7. The diffraction coefficients.

7.1. The diffraction coefficients for bulk waves. Let us first define the function

$$E(k_i^{(j)}r) = \sqrt{\frac{i}{8\pi}}\frac{e^{ik_i^{(j)}r}}{\sqrt{k_i^{(j)}r}}e^{-ik_i^{(j)}y_1p_1}, \tag{7.1}$$

where $\mathbf{p} = (\cos\theta, \sin\theta)$ is a unit vector in the direction of the diffracted wave, with the angle θ such that $\cos\theta = x_1/r$ and r is the distance to the origin. Let $\mathbf{p}^\perp$ be another unit vector, which is orthogonal to $\mathbf{p}$. It then can be chosen to be $\mathbf{p}^\perp = (-p_2, p_1) = (-\sin\theta, \cos\theta)$. It is easy to check that as $k_i^{(j)}r \to \infty$, the leading term in the expansion of the argument gives

$$\begin{aligned} H_0\left(k_i^{(j)}\sqrt{(x_1-y_1)^2+x_2^2}\right) &= H_0\left(k_i^{(j)}\left[r - y_1\cos\theta + O\left(\frac{1}{r}\right)\right]\right) \\ &\approx H_0(k_i^{(j)}r) \approx -4iE(k_i^{(j)}r). \end{aligned} \tag{7.2}$$

In the far field, the Green's tensor and Green's stress can be decomposed as

$$\begin{aligned} u_{ik}^{G(j)}(x_1-y_1,x_2) &= u_{ik}^{G1(j)}(x_1-y_1,x_2) + u_{ik}^{G2(j)}(x_1-y_1,x_2), \\ \sigma_{2ik}^{G(j)}(x_1-y_1,x_2) &= \sigma_{2ik}^{G1(j)}(x_1-y_1,x_2) + \sigma_{2ik}^{G2(j)}(x_1-y_1,x_2), \end{aligned} \tag{7.3}$$

where the superscript $i(j)$ refers to longitudinal ($i = 1$) or transverse ($i = 2$) wave in the medium $I^{(j)}$, and in the far field, as $k_i^{(j)}r \to \infty$, we can use approximations

$$\begin{aligned} u_{ik}^{G1(j)}(x_1-y_1,x_2) &\approx \left[\frac{1}{\mu^{(j)}}(\kappa^{(j)})^{-2}p_iE(k_1^{(j)}r)\right]p_k, \\ \sigma_{2ik}^{G1(j)}(x_1-y_1,x_2) &\approx ik_1^{(j)}\left[(\delta_{i2} - 2(\kappa^{(j)})^{-2}p_1p_i^\perp)E(k_1^{(j)}r)\right]p_k, \\ u_{ik}^{G2(j)}(x_1-y_1,x_2) &\approx \left[\frac{1}{\mu^{(j)}}p_i^\perp E(k_2^{(j)}r)\right]p_k^\perp, \\ \sigma_{2ik}^{G2(j)}(x_1-y_1,x_2) &\approx ik_2^{(j)}\left[(p_2p_i^\perp + p_1p_i)E(k_2^{(j)}r)\right]p_k^\perp \end{aligned} \tag{7.4}$$

(see [11], [21], or Appendix A).

To continue, for the scattered field $\mathbf{u}^{sc(j)}(\mathbf{x})$, the integral representation (2.1) can be rewritten as

$$\begin{aligned} H[(-1)^{j+1}x_2]u_k^{sc(j)}(\mathbf{x}) = (-1)^j\sum_{i=1}^{2}\int_{-\infty}^{\infty}&[\sigma_{2ik}^{G(j)}(x_1-y_1,x_2)u_i^{sc(j)}(y_1,0) \\ &+ u_{ik}^{G(j)}(x_1-y_1,x_2)\sigma_{2i}^{(j)sc}(y_1,0)]dy_1 \end{aligned} \tag{7.5}$$

(see, e.g., [11]). Substituting (7.4) into (7.3) and the result into the version of (2.1) applicable to the scattered field, for $0 < \theta < 2\pi$ we can write

$$H(\pi-\theta)u_k^{tip(j)} \approx D^{1(j)}(\theta)\frac{e^{ik_1^{(j)}r}}{\sqrt{k_1^{(j)}r}}p_k + D^{2(j)}(\theta)\frac{e^{ik_2^{(j)}r}}{\sqrt{k_2^{(j)}r}}p_k^\perp,$$

where the diffraction coefficients for diffracted body waves can be expressed in terms of the displacement vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, and traction vector $\widehat{\mathbf{t}}(\xi)$ in the following manner:

$$\begin{aligned}
D^{1(1)}(\theta) &= -\sqrt{\frac{-i}{8\pi}}(\kappa^{(1)})^{-2}\Big\{2p_1p_2\widehat{u}_1^{(1)}(-p_1)+[(\kappa^{(1)})^2-2p_1^2]\widehat{u}_2^{(1)}(-p_1)\\
&\qquad\qquad +p_1\widehat{t}_1(-p_1)+p_2\widehat{t}_2(-p_1)\Big\},\\
D^{1(2)}(\theta) &= \sqrt{\frac{-i}{8\pi}}\Big[\kappa_1^{(2)}\{2(\kappa^{(2)})^{-2}p_1p_2\widehat{u}_1^{(2)}(-\kappa_1^{(2)}p_1)+[1-2(\kappa^{(2)})^{-2}p_1^2]\widehat{u}_2^{(2)}(-\kappa_1^{(2)}p_1)\}\\
&\qquad +\mu(\kappa^{(2)})^{-2}\{p_1\widehat{t}_1(-\kappa_1^{(2)}p_1)+p_2\widehat{t}_2(-\kappa_1^{(2)}p_1)\}\Big],\\
D^{2(1)}(\theta) &= -\sqrt{\frac{-i}{8\pi}}\Big[\kappa^{(1)}\{[p_1^2-p_2^2]\widehat{u}_1^{(1)}(-\kappa_2^{(1)}p_1)+2p_1p_2\widehat{u}_2^{(1)}(-\kappa_2^{(1)}p_1)\}\\
&\qquad -p_2\widehat{v}_3^-(-\kappa_2^{(1)}p_1)+p_1\widehat{v}_4^-(-\kappa_2^{(1)}p_1)\Big],\\
D^{2(2)}(\theta) &= \sqrt{\frac{-i}{8\pi}}\Big[\kappa_2^{(2)}\{[p_1^2-p_2^2]\widehat{u}_1^{(2)}(-\kappa_2^{(2)}p_1)+2p_1p_2\widehat{u}_2^{(2)}(-\kappa_2^{(2)}p_1)\}\\
&\qquad +\mu\{-p_2\widehat{v}_3^-(-\kappa_2^{(2)}p_1)+p_1\widehat{v}_4^-(-\kappa_2^{(2)}p_1)\}\Big],
\end{aligned}\tag{7.6}$$

where if $\widehat{\mathbf{s}}(\xi)$ is known, $\widehat{\mathbf{t}}(\xi)$ can be found using (4.5). An additional difficulty is presented by the fact that, for $\xi>1$, $\widehat{\mathbf{s}}(\xi)$ as defined in (4.10) contains a singular integral (5.3). Since in (7.6) the arguments of all $\widehat{\mathbf{t}}$ components are given in the form $\xi=-\kappa_i^{(j)}p_1$, the function that needs evaluating is $\widehat{\mathbf{s}}(-\kappa_i^{(j)}p_1)$. When $p_1\geq 0$ the evaluation can be carried out using (4.10), and when $p_1\leq 0$ the combination of (4.7) and (4.6) should be used instead. To calculate the displacement vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, we note that the system (3.11) can be rewritten in terms of $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, and $\widehat{\mathbf{t}}(\xi)$ as

$$p(\xi)\mathbf{v}^{(j)inc}=\widehat{A}^{(j)+}(\xi)\widehat{\mathbf{u}}^{(j)}(\xi)+\widehat{A}^{(j)-}(\xi)\widehat{\mathbf{t}}(\xi),\quad j=1,2,\tag{7.7}$$

where $\widehat{A}^{(j)\pm}(\xi)$ are 2×2 matrices, which form matrices $\widehat{A}^{\pm}(\xi)$ in (3.11), so that we have

$$\widehat{A}^{+}(\xi)=\begin{pmatrix}\widehat{A}^{(1)+}(\xi) & 0\\ 0 & \widehat{A}^{(2)+}(\xi)\end{pmatrix},\qquad A^{-}(\xi)=\begin{pmatrix}\widehat{A}^{(1)+}(\xi) & \widehat{A}^{(1)-}(\xi)\\ \widehat{A}^{(2)+}(\xi) & \widehat{A}^{(2)-}(\xi)\end{pmatrix},\tag{7.8}$$

and vectors $\mathbf{v}^{(j)inc}(\xi)$ are

$$\mathbf{v}^{(1)inc}(\xi)=\begin{pmatrix}v_1^{inc}(\xi)\\ v_2^{inc}(\xi)\end{pmatrix},\quad \mathbf{v}^{(2)inc}(\xi)=\begin{pmatrix}v_3^{inc}(\xi)\\ v_4^{inc}(\xi)\end{pmatrix}.\tag{7.9}$$

Therefore, after $\widehat{\mathbf{t}}(\xi)$ is found, vectors $\widehat{\mathbf{u}}^{(j)}(\xi),\ j=1,2$, can be calculated using (7.7). Note that since $|\kappa_i^{(j)}p_1|\leq\max\{\kappa_i^{(j)}\}<\xi^{R(j)}$, matrices $\widehat{A}^{(j)+}(-\kappa_i^{(j)}p_1)$ are regular.

7.2. The Rayleigh diffraction coefficients. On each of the traction-free surfaces $(x_2=0,x_1>0)$ the Rayleigh wave can be defined as

$$\mathbf{u}^{R(j)}(y_1,0)=D^{(j)R}\mathbf{v}^{(j)R}e^{ik^{R(j)}y_1},\qquad y_1>0,\tag{7.10}$$

where the so-called Rayleigh diffraction coefficient $D^{(j)R}$ is its amplitude, $k^{R(j)} = k_1^{(1)}\xi^{R(j)}$, and the unit vector $\mathbf{v}^{(j)R}$ is a nonzero solution of the homogeneous equation

$$\widehat{A}^{+}(-\xi^{R(j)})\mathbf{v}^{(j)R} = \mathbf{0}. \tag{7.11}$$

It can be shown that the unit vector $\mathbf{v}^{(j)R}$ is

$$\mathbf{v}^{(j)R} = \left(-\frac{2\xi^{R(j)}\gamma_2^{(j)}(\xi^{R(j)})}{[\kappa_2^{(j)}]^2}, (-1)^{j+1}\frac{a_j(\xi^{R(j)})}{[\kappa_2^{(j)}]^2}\right)^T. \tag{7.12}$$

Applying the Fourier transform to (7.10), we obtain

$$\widehat{\mathbf{u}}^{R(j)}(\xi) = \frac{iD^{(j)R}\mathbf{v}^{(j)R}}{\xi + \xi^{R(j)}}. \tag{7.13}$$

Multiplying (7.7) by $[\widehat{A}^{(j)+}(\xi)]^{-1}$ gives us

$$p(\xi)[\widehat{A}^{(j)+}(\xi)]^{-1}\mathbf{v}^{(j)inc} = \widehat{\mathbf{u}}^{(j)}(\xi) + \frac{1}{R_j}\widehat{B}^{(j)-}(\xi)\widehat{\mathbf{t}}(\xi), \quad j = 1, 2, \tag{7.14}$$

with a finite matrix

$$\widehat{B}^{(j)-}(\xi) = \mu_j \begin{pmatrix} (-1)^j b_{j1}(\xi) & -g_j(\xi) \\ g_j(\xi) & (-1)^j b_{j2}(\xi) \end{pmatrix}, \quad j = 1, 2. \tag{7.15}$$

Let us now evaluate the residue of both sides of (7.14) at $\xi = -\xi^{R(j)}$. This can be done by multiplying them by $\xi + \xi^{R(j)}$ and finding the limits when $\xi \to -\xi^{R(j)}$. By definition of $p(\xi)$, the left-hand side of the resulting equation is zero. The residue of the displacement vector at the Rayleigh pole $\xi = -\xi^{R(j)}$ is

$$\operatorname*{Res}_{\xi=-\xi^{R(j)}} \widehat{\mathbf{u}}^{(j)}(\xi) = \lim_{\xi\to-\xi^{R(j)}} (\xi + \xi^{R(j)})\widehat{\mathbf{u}}^{R(j)}(\xi) = iD^{(j)R}\mathbf{v}^{(j)R}, \quad j = 1, 2. \tag{7.16}$$

Therefore, (7.14) leads to

$$iD^{(j)R}\mathbf{v}^{(j)R} + \frac{1}{R_j'(-\xi^{R(j)})}\widehat{B}^{(j)-}(-\xi^{R(j)})\widehat{\mathbf{t}}(-\xi^{R(j)}) = \mathbf{0}, \quad j = 1, 2, \tag{7.17}$$

where $R_j'(\xi)$ is the derivative of the Rayleigh function $R_j(\xi)$. It follows that, for each medium $I^{(j)}$, $j = 1, 2$, the respective Rayleigh diffraction coefficient can be expressed via $\widehat{\mathbf{t}}(\xi)$ as

$$D^{(j)R} = \frac{i[\widehat{B}_{11}^{(j)-}(-\xi^{R(j)})\widehat{t}_1(-\xi^{R(j)}) + \widehat{B}_{12}^{(j)-}(-\xi^{R(j)})\widehat{t}_2(-\xi^{R(j)})]}{R_j'(-\xi^{R(j)})v_1^{(j)R}}, \quad j = 1, 2. \tag{7.18}$$

8. Numerical results. We have developed a FORTRAN90 program for computing the diffraction coefficient $D^{i(j)}(\theta)$ using (7.6), where the displacements and tractions are evaluated at $\xi = -\kappa_i^{(j)}\cos\theta$, with θ being an observation angle.

As has been discussed above, $\widehat{\mathbf{u}}^{(j)}(\xi)$ and $\widehat{\mathbf{t}}(\xi)$ are both singular at the GE pole $\xi = \xi^{inc}$. At the real angles $\theta = \theta^{sh}$, which satisfy the equation

$$\kappa_i^{(j)}\cos(\theta^{sh}) = -\kappa_n^{(m)}\cos(\theta^{inc}), \tag{8.1}$$

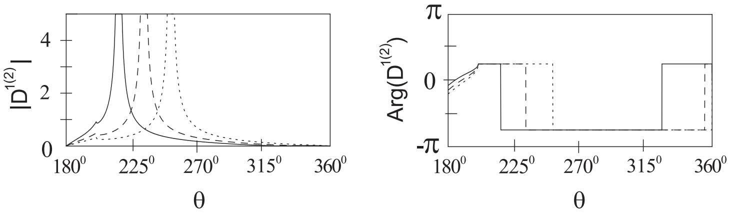


FIG. 8.1. *The longitudinal diffraction coefficient in the lower medium. The solid line corresponds to the angle of incidence of $\theta^{inc} = 30^o$, the dashed line to 50^o, and the dotted line to 70^o.*

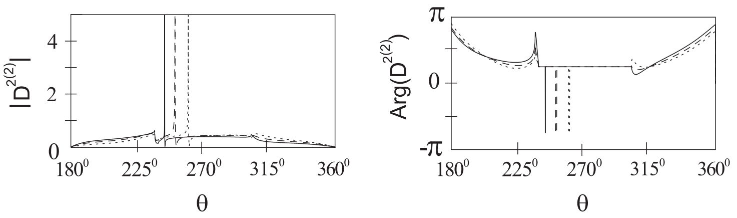


FIG. 8.2. *The transverse diffraction coefficient in the lower medium. The key is as above.*

the estimates of $|D^{i(j)}(\theta)|$ (incorrect near these angles) are infinite and the estimates of $\mathrm{Arg}[D^{i(j)}(\theta^{sh})]$ experience a π jump. When $|\kappa_n^{(m)}\cos(\theta^{inc})/\kappa_i^{(j)}| > 1$, (8.1) has no real valued solutions and $D^{i(j)}(\theta)$ are correct estimates, continuous at all observation angles. The solution of (8.1) is

$$\theta^{sh} = \pi + (-1)^j \arccos\left(\frac{\kappa_n^{(m)}}{\kappa_i^{(j)}}\cos(\theta^{inc})\right). \tag{8.2}$$

When (8.2) defines a real valued angle, it is the shadow boundary of either the reflected or refracted wave in the incident medium, or either the transmitted longitudinal or transverse wave in the other medium.

Other special angles, which can be seen on the graphs in this section, are the so-called critical angles θ^{cr}. They describe the boundaries of the regions that support head waves and correspond to the branch points, that is, satisfy the equation

$$\pm\kappa_k^{(l)} = -\kappa_i^{(j)}\cos\theta^{cr}, \qquad l, k = 1, 2. \tag{8.3}$$

These critical angles do not depend on the angle of incidence. Again, the approximation method used in section 7.1 fails in their vicinity, and they show up on the graphs as small blips.

As an illustration, Figures 8.1–8.4 present diffraction coefficients for a semi-infinite crack, which is sandwiched between aluminum and steel.

The incident plane wave is a longitudinal wave, incoming from the aluminum half plane. The amplitudes of the diffraction coefficients are presented on the left and phases on the right. The model parameters are as follows. In medium 1 (aluminum), density $\rho^{(1)} = 2700$ kg/m^3, longitudinal speed $c_1^{(1)} = 6300$ m/s, and shear speed $c_2^{(1)} = 3100$ m/s. In medium 2 (steel), density $\rho^{(2)} = 7800$ kg/m^3, longitudinal speed $c_1^{(2)} = 5900$ m/s, and shear speed $c_2^{(2)} = 3200$ m/s. We can see the geometrical shadow boundaries described by (8.2), where the amplitude of the formally evaluated

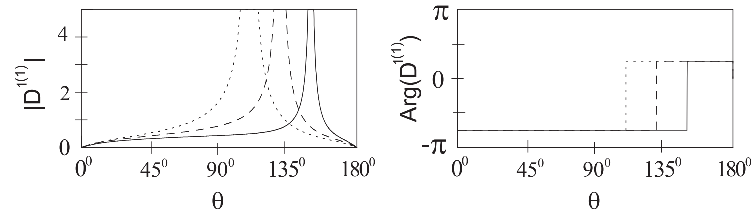


FIG. 8.3. *The longitudinal diffraction coefficient in the upper medium. The key is as in Figure* 8.1.

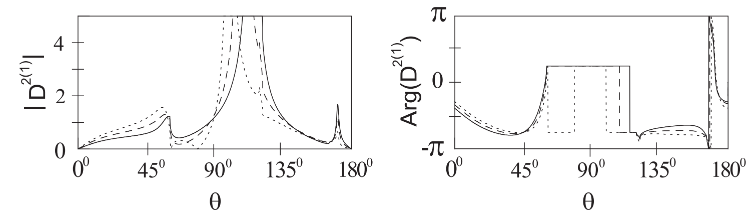


FIG. 8.4. *The transverse diffraction coefficient in the upper medium. The key is as in Figure* 8.1.

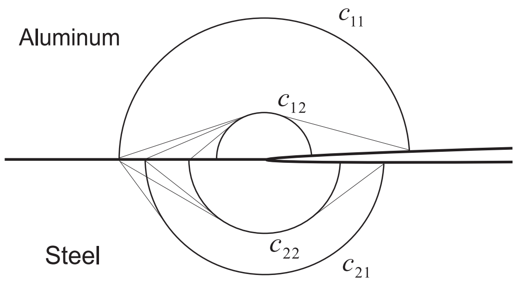


FIG. 8.5. *Head wave fronts.*

diffraction coefficient is infinite and the phase experiences a π jump. Since the critical angles do not depend on the angle of incidence, the corresponding blips appear at the same place for all three curves. The corresponding head wave fronts are shown in Figure 8.5. It can be seen that in the upper medium, the diffracted longitudinal wave is not affected by head waves. This is due to the fact that in aluminum the longitudinal speed is greater than in steel.

In Figure 8.1 the critical angle is $\theta \approx 200^o$. The head wave affects $D^{2(2)}$ at $\theta \approx 237.2^o$, $\theta \approx 239.5^o$, and $\theta \approx 328^o$. In Figure 8.2 only two blips are seen at about $\theta \approx 237^o$ and $\theta \approx 303.6^o$. This is due to the small difference between the first two critical angles $\theta \approx 237.2^o$ and $\theta \approx 239.5^o$. By decreasing both the discretization step and the interval of observation angles, the critical angles separate. The head wave affects $D^{2(1)}$ at $\theta \approx 60.5^o$, $\theta \approx 119.5$, $\theta \approx 121.7^o$, and $\theta \approx 165.6^o$. Again, in Figure 8.4 the critical angles $\theta \approx 119.5^o$ and $\theta \approx 121.7^o$ lie too close to each other. The overall conclusion is that for, say, the 30^o incidence, the longitudinal tip diffracted waves, which propagate in the upper medium, can be best detected at observation angles between 40^o and 130^o. As the angle of incidence increases, the range of advantageous observation angles shortens (see Figure 8.3).

The Rayleigh diffraction coefficients do not depend on the observation angle and are calculated using (7.18). Note that the Rayleigh speeds for aluminum and steel are $c_{Al} = 2894$ m/s and $c_{St} = 2964$ m/s, respectively, and the bimaterial constant is $\nu_0 = 0.0330434$. For the angle of incidence $\theta^{inc} = 30^{\text{o}}$, in aluminum $D^{(1)R} = -0.1208 - 0.206i$, and in steel $D^{(2)R} = 0.1428{-}0.2128i$; for $\theta^{inc} = 50^{\text{o}}$, $D^{(1)R} = -0.1004{-}0.3094i$ and $D^{(2)R} = 0.1285 - 0.3272i$; and for $\theta^{inc} = 70^{\text{o}}$, $D^{(1)R} = -0.0517 - 0.4005i$ and $D^{(2)R} = 0.0784 - 0.4285i$.

We finish this section by discussing the code testing. As already mentioned, the critical angles and positions of the shadow boundaries of reflected and refracted waves, when calculated independently, agree with the above graphs. We also know the phase (either $\pi/4$ or $-3\pi/4$) on those portions of the wave front which are not occupied by the head waves—these are correct. Another stringent internal test is to evaluate the left-hand side of (4.6) on the interval $\xi \,\epsilon\, [-1, 1]$. The right-hand side of this equation is known. The left-hand side has been computed using our code. The maximum relative error of each component has been found to be 2%, which is satisfactory. Finally, one could make another check by considering the limiting case of the two identical half planes: In this case, the off-diagonal terms in (4.4) become zero, and (4.3) reduces to the equivalent of the decoupled Wiener–Hopf equations for the crack opening displacements as found in the studies of the isotropic case, e.g., [11]. However, no numerical check of this nature has been conducted, because, when both half planes are the same, (4.7) contains a dipole instead of a pole, and although one could use this representation, it would require additional careful programming: In (4.7) one would have to use the sum of the two poles instead of the product. Even then, numerical difficulties would still arise for nearly identical top and bottom media.

9. Conclusions.

We have developed a semianalytical approach to calculating diffraction coefficients for a semi-infinite crack sandwiched between two different isotropic media. We have introduced a stable numerical scheme for solving the resulting system of integral equations, (5.7). Our main achievement has been to produce a fast computer code, which is applicable to any pair of (sufficiently different) isotropic materials which do not support the Stoneley wave and are irradiated by a plane wave incident from either medium. The incident wave can be longitudinal or transverse and incoming at an arbitrary angle. The absence of the Stoneley wave does not constitute a serious restriction, since this case is ubiquitous in applications. Nevertheless, we plan to publish another paper modeling materials where the Stoneley wave is present too. As an illustration, we have presented plots of diffraction coefficients for a crack sandwiched between aluminum and steel.

Appendix A. The two dimensional Green's tensor and Green's stress tensor.

Since the incident wave can be considered as radiated by a line load, both the two dimensional Green's tensor and Green's stress tensor can be represented in terms of the Hankel function of the first kind of the zeroth order, $H_0 \equiv H_0^{(1)}$, and its derivatives (see [11], [21]), so that at any observation point $\mathbf{x}$ we have

$$-4i\mu^{(j)} u_{ik}^{G(j)}(\mathbf{x}) = \frac{1}{[k_2^{(j)}]^2}\left[-H_0(k_1^{(j)}r) + H_0(k_2^{(j)}r)\right]_{,ik} + H_0(k_2^{(j)}r)\delta_{ik},$$

$$\begin{aligned}-4i\sigma_{2ik}^{G(j)}(\mathbf{x}) &= \left\{1 - 2\left[\frac{c_2^{(j)}}{c_1^{(j)}}\right]^2\right\}\left[H_0(k_1^{(j)}r)\right]_{,k}\delta_{2i} - \frac{2}{(k_2^{(j)})^2}\left[H_0(k_1^{(j)}r) - H_0(k_2^{(j)}r)\right]_{,2ik} \\ &\quad + \left[H_0(k_2^{(j)}r)\right]_{,2}\delta_{ik} + \left[H_0(k_2^{(j)}r)\right]_{,i}\delta_{2k}, \qquad \text{(A.1)}\end{aligned}$$

where $\mu^{(j)},\ j=1,2$, is the shear modulus in medium $I^{(j)}$; r is the distance to the origin; and an index, say k, after the comma refers to differentiation with respect to the corresponding spatial variable x_k. Applying operations of dilatation and rotation to both sides of (A.1) gives us

$$\begin{aligned}
4i\mu^{(j)}u_i^{G(j)1}(\mathbf{x}) &= \left[\frac{c_2^{(j)}}{c_1^{(j)}}\right]^2 \left[H_0(k_1^{(j)}r)\right]_{,i}\,,\\
-4i\sigma_i^{G(j)1}(\mathbf{x}) &= (k_1^{(j)})^2\left\{\left[2\left(\frac{c_2^{(j)}}{c_1^{(j)}}\right)^2-1\right]H_0(k_1^{(j)}r)\delta_{2i}+\frac{2}{(k_2^{(j)})^2}\left[H_0(k_1^{(j)}r)\right]_{,2i}\right\},\\
-4i\mu^{(j)}u_i^{G(j)2}(\mathbf{x}) &= \left[H_0(k_2^{(j)}r)\right]_{,1}\delta_{2i}-\left[H_0(k_2^{(j)}r)\right]_{,2}\delta_{1i},\\
-4i\sigma_i^{G(j)2}(\mathbf{x}) &= \left[H_0(k_2^{(j)}r)\right]_{,12}\delta_{2i}-\left[H_0(k_2^{(j)}r)\right]_{,22}\delta_{1i}+\left[H_0(k_2^{(j)}r)\right]_{,1i}\,.
\end{aligned}\tag{A.2}$$

Appendix B. The extinction theorem. The extinction theorems are easily proved for finite sources and obstacles using Green's theorem. Difficulties arise when the incident waves are plane and obstacles infinite. One approach to dealing with this complication is to develop methods such as those offered in [22] and references therein (also see [23]). Below we offer an alternative justification.

Let us focus on the scattered field in the upper plane. Any identity involving the incident field can be established by direct integration. For simplicity of presentation, we omit the superscript $^{(1)}$. Then solving the Fourier transform of the equations of motion for the elastic solid gives

$$\widehat{\mathbf{u}}^{sc}(\xi,x_2)=A(\xi)(-\xi,\gamma_1)^Te^{ik_1\gamma_1x_2}+B(\xi)(\gamma_2,\xi)^T\mathrm{e}^{ik_1\gamma_2x_2},\quad x_2>0,\tag{B.1}$$

where $A(\xi)$ and $B(\xi)$ are unknown. The solutions proportional to $\exp[-ik_1\gamma_ix_2]$, $i=1,2$, are rejected because they do not satisfy the radiation conditions: Either they are incoming from infinity or else, when we move the branches off the real axis (see Figure C.1 below) as $x_2\to\infty$, they become unbounded. It follows that on the top face of the crack, $x_2=0^+$, we have

$$\widehat{\mathbf{u}}^{sc}(\xi,0^+)=A(\xi)(-\xi,\gamma_1)^T+B(\xi)(\gamma_2,\xi)^T.\tag{B.2}$$

It can easily be verified that a similar formula holds for the traction related vector $\widehat{\mathbf{t}}^{sc}$ (see (4.1)),

$$\widehat{\mathbf{t}}^{sc}(\xi,0^+)=-i[A(\xi)(2\xi\gamma_1,2\xi^2-\kappa_2^2)^T+B(\xi)(2\xi^2-\kappa_2^2,-2\xi\gamma_2)^T].\tag{B.3}$$

The solution to the Fourier transform of the equations of motion, which is valid in both half planes, is

$$\widehat{\mathbf{u}}^{sc}(\xi,x_2)=A_1(\xi)(-\xi,\gamma_1\mathrm{sgn}(x_2))^Te^{ik_1\gamma_1|x_2|}+B_1(\xi)(\gamma_2\mathrm{sgn}(x_2),\xi)^T\mathrm{e}^{ik_1\gamma_2|x_2|},\ x_2>0,\tag{B.4}$$

where we have

$$\begin{aligned}
2\kappa_2^2\gamma_1A_1(\xi) &= -2\xi\gamma_1\mathrm{sgn}(x_2)\widehat{u}_1^{sc}(\xi,0^+)+(\kappa_2^2-2\xi^2)\widehat{u}_2^{sc}(\xi,0^+)\\
&\quad+i[\xi\widehat{t}_1^{sc}(\xi,0^+)-\gamma_1\mathrm{sgn}(x_2)\widehat{t}_2^{sc}(\xi,0^+)],\\
2\kappa_2^2\gamma_2B_1(\xi) &= (\kappa_2^2-2\xi^2)\widehat{u}_1^{sc}(\xi,0^+)+2\xi\gamma_2\mathrm{sgn}(x_2)\widehat{u}_2^{sc}(\xi,0^+)\\
&\quad-i[\gamma_2\mathrm{sgn}(x_2)\widehat{t}_1^{sc}(\xi,0^+)+\xi\widehat{t}_2^{sc}(\xi,0^+)].
\end{aligned}\tag{B.5}$$

Substituting (B.2) and (B.3) into (B.5) yields

$$(A_1(\xi), B_1(\xi)) = (A(\xi), B(\xi))H(x_2). \tag{B.6}$$

Thus, the scattered field defined by (B.4) agrees with (B.1) in the upper half plane and is identically zero in the fictitious half plane $x_2 < 0$. The inverse transform of (B.4) leads to the extinction theorem (2.1) for the scattered field.

Appendix C. Singularities in (3.11). Let us describe all singularities of functions that appear in (3.11). First, in view of (3.10), the left-hand side of (3.11) has simple poles at $\xi = \xi^{inc} + i0$ and $\xi = \xi^{inc} - i0$, which correspond to the incident and reflected bulk waves, respectively. They give rise to GE bulk waves.

Second, it is easy to check that the determinant of the matrix $\widehat{A}^{+}(\xi)$, $|\widehat{A}^{+}(\xi)|$ is a product of two Rayleigh functions, $R_1(\xi)$ and $R_2(\xi)$,

$$R_j(\xi) = a_j^2(\xi) + 4\xi^2\gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi), \tag{C.1}$$

where the subscript $j = 1, 2$ refers to medium $I^{(j)}$ (see, e.g., [21]). Thus, the solutions $\xi = \pm\xi^{R(j)}$ of the equation $|\widehat{A}^{+}(\xi)| = 0$ are zeros of $R_1(\xi)$ and $R_2(\xi)$ and can be shown to be simple (distinct). The zeros $\xi = -\xi^{R(j)}$ are known to give rise to the outgoing Rayleigh surface waves.

It is equally easy to check that $|\widehat{A}^{-}(\xi)|$ is the well-known Stoneley function

$$S(\xi) = \mu^2 R_1(\xi)h_2(\xi) + R_2(\xi)h_1(\xi) + \mu[b_{11}(\xi)b_{22}(\xi) + b_{21}(\xi)b_{12}(\xi) - 2g_1(\xi)g_2(\xi)] \tag{C.2}$$

(see, e.g., [24]), where we use the notation

$$\begin{aligned} &b_{j1}(\xi) = [\kappa_2^{(j)}]^2\gamma_2^{(j)}(\xi), \qquad b_{j2}(\xi) = [\kappa_2^{(j)}]^2\gamma_1^{(j)}(\xi), \\ &g_j(\xi) = \xi[2\gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi) - a_j(\xi)], \qquad h_j(\xi) = \gamma_1^{(j)}(\xi)\gamma_2^{(j)}(\xi) + \xi^2, \quad j = 1, 2. \end{aligned} \tag{C.3}$$

In general, the zero of $S(\xi) = 0$ (which is also simple) can give rise to an outgoing Stoneley wave. Using Cagniard's method (see [24]) we have established that for the set of parameters used in this paper such a solution does not exist, and therefore no Stoneley surface wave runs between materials under study. We remark in passing that this situation is common, and Cagniard [24] refers to the Stoneley wave as "a rather special phenomenon," meaning that it exists only in narrow ranges of material parameters.

To continue, both matrices $\widehat{A}^{\pm}(\xi)$ involve multivalued radicals $\gamma_i^{(j)}(\xi)$ defined below (3.14). In order to render the matrices single valued we introduce the branch cuts $C_i^{(j)\mp}$, $i, j = 1, 2$, which run between branch points $\pm\kappa_i^{(j)}$, defined below (3.9), and $\pm\infty$, respectively. Let us apply the limiting absorption principle and replace $\kappa_i^{(j)}$ by $\kappa_i^{(j)} + i\epsilon_1$, $\epsilon_1 > 0$. This shifts the branch cuts away from the real axis as indicated in Figure C.1, and when performing the inverse Fourier transform, the corresponding singularities give rise to the waves, which satisfy the radiation condition, that is, are outgoing to infinity.

The radicals $\gamma_i^{(j)}(\xi)$ can be factorized so that we have

$$\gamma_i^{(j)}(\xi) = \gamma_i^{(j)+}(\xi)\gamma_i^{(j)-}(\xi), \tag{C.4}$$

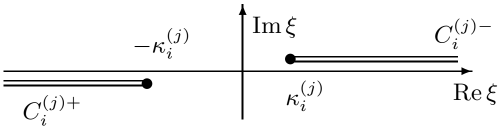


FIG. C.1. *The branch cuts of* $\gamma_i^{(j)}(\xi)$.

where we use the notation

$$\gamma_i^{(j)+}(\xi) = \sqrt{\kappa_i^{(j)} + \xi}, \quad \gamma_i^{(j)-}(\xi) = \gamma_i^{(j)+}(-\xi), \tag{C.5}$$

and only the cut $C_i^{(j)+}$ is required to render $\gamma_i^{(j)+}(\xi)$ single-valued. The radical $\gamma_i^{(j)-}(\xi)$ is rendered single-valued by $C_i^{(j)-}$. Note that if $\xi = -\zeta$ lies on the branch cut $C_i^{(j)+}$, the definition implies that ξ lies on the upper side of the cut, so that $\xi = -\zeta + i0$. It follows that $\gamma_i^{(j)+}(-\zeta)$ is well defined. As ξ tends to $-\zeta$ from below the branch cut, we have

$$\gamma_i^{(j)+}(\xi) \to -\gamma_i^{(j)+}(-\zeta) \equiv \gamma_i^{(j)+}(-\zeta - i0). \tag{C.6}$$

Note that in the main text we drop a combination of subscript and superscript ${}_1^{(1)}$ in the symbol for the longest branch cut $C_1^{(1)\pm}$. To summarize, the known functions in (3.11) involve two GE poles $\xi^{inc} \pm i0$, two Rayleigh poles $-\xi^{R(1)}$ and $-\xi^{R(2)}$, as well as two branch cuts $C^{\pm}$.

Appendix D. Auxiliary functions and vectors. Let us now describe auxiliary functions used in the main text. Let four branch points be sorted in order of the descending moduli, with $\kappa_1 = \min\{\kappa_i^{(j)}\}$ and $\kappa_4 = \max\{\kappa_i^{(j)}\}$, $i, j = 1, 2$, denoting the corresponding radicals $\gamma_i^{\pm}(\xi) = \sqrt{\kappa_i \pm \xi}$ and the respective branch cuts $C_i^+ = \{\xi : \xi \leq -\kappa_i\}$ and $C_i^- = \{\xi : \xi \geq \kappa_i\}$, $i = 1, 2, 3, 4$.

For each pair of numbers $a \neq -\kappa_i$ and ξ, let us introduce the auxiliary functions $H_i^{\pm}(\xi, a)$,

$$H_i^+(\xi, a) = \frac{\gamma_i^+(a) - \gamma_i^+(\xi)}{2\gamma_i^+(a)} \left[\frac{a_0 - \gamma_i^+(a)}{a_0 + \gamma_i^+(\xi)}\right]^{2n+1}, \qquad H_i^-(\xi, a) = H_i^+(-\xi, -a), \tag{D.1}$$

where $n \geq 3$ and a_0 lies far away from the branch cut C_i^+. To be specific, let $a_0 = 1+i$. Then the function H_i^+ has the following properties:

- It has no poles in ξ.
- The branch cut C^+ renders it single-valued.
- For any $\xi \in C_i^+$ we have

  $$H_i^+(\xi + i0, \xi) = 0 \quad \text{and} \quad H_i^+(\xi - i0, \xi) = 1, \tag{D.2}$$

  where, as the above notation suggests, $H_i^+(\xi + i0, a)$ and $H_i^+(\xi - i0, a)$ are values of $H_i^+(\xi, a)$ evaluated on the upper and lower sides of C_i^+, respectively. (We recall that, unless stated otherwise, ξ lies on the positive side of the cut.)
- For any $\xi \in C_i^-$ such that Re $\zeta \gg 1$ we have

  $$H_i^+(-\xi + i0, a) - H_i^+(-\xi - i0, a) \sim \frac{\text{constant}}{[\gamma_i^+(-\xi)]^{2n}}. \tag{D.3}$$

Let us now define branch cuts

$$\begin{aligned} &C^-_{ii+1} = \{\xi : \kappa_i \le \xi \le \kappa_{i+1}\}, \quad i = 1, 2, 3,\\ &C^-_{45} = \{\xi : \xi \ge \kappa_4\},\\ &C^+_{ii+1} = \{\xi : -\kappa_{i+1} \le \xi \le -\kappa_i\}, \quad i = 1, 2, 3,\\ &C^+_{45} = \{\xi : \xi \le -\kappa_4\}, \end{aligned} \tag{D.4}$$

and introduce the auxiliary functions

$$\begin{aligned} &F^{\pm}_i(\xi, a) = H^{\pm}_i(\xi, a) - H^{\pm}_{i+1}(\xi, a), \quad i = 1, 2, 3,\\ &F^{\pm}_4(\xi, a) = H^{\pm}_4(\xi, a). \end{aligned} \tag{D.5}$$

It is easy to see that each function $F^{\pm}_i(\xi, a)$ has a branch cut $C^{\pm}_{ii+1}$, $i = 1, 2, 3, 4$, and $F^{\pm}_i(\xi, \xi) = 1$ on the negative side of its branch cut and 0 everywhere else. Let us introduce a function $G(\xi)$ as

$$G(\xi) = \left(\sqrt{1+\xi} + \sqrt{\kappa^{(1)}_2 + \xi}\,\right)^2, \tag{D.6}$$

which is real outside the interval $(-\kappa^{(1)}_2, -1)$ and $O(\xi)$ at infinity.

Appendix E. Auxiliary relationships. Let us determine eight scalar constants $D^{\pm}_i$, $i = 1, 2, 3, 4$, introduced in (6.1). Let us show that they are linearly dependent, and therefore that the total number of unknowns can be decreased by four. Let us do this by analyzing the asymptotic behavior of both sides of (6.2). As $\xi \to \infty$, matrix $M(\xi) \to M_\infty$,

$$M_\infty = \begin{pmatrix} 0_2 & m^-_\infty I_2 \\ m^+_\infty I_2 & 0_2 \end{pmatrix}, \tag{E.1}$$

where matrices 0_2 and I_2 denote the zero and identity 2×2 matrices, respectively; $m^{\pm}_\infty$ are known constants,

$$\begin{aligned} m^-_\infty &= -\frac{1}{8}\left\{ \frac{[\kappa^{(1)}_2]^2}{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2} + \mu \frac{[\kappa^{(2)}_2]^2}{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2} \right\},\\ m^+_\infty &= -8[S_\infty]^{-1}\left([\kappa^{(1)}_2]^2\{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2\} + \mu[\kappa^{(2)}_2]^2\{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2\} \right); \end{aligned} \tag{E.2}$$

and we use the notation $S_\infty = \lim_{\xi\to\infty} S(\xi)/\xi^2$. Using the Stoneley function $S(\xi)$ defined in (C.2), we find

$$\begin{aligned} S_\infty = \mu^2&\{[\kappa^{(1)}_2]^2 - [\kappa^{(1)}_1]^2\}\{[\kappa^{(2)}_2]^2 + [\kappa^{(2)}_1]^2\} + \{[\kappa^{(1)}_2]^2 + [\kappa^{(1)}_1]^2\}\{[\kappa^{(2)}_2]^2 - [\kappa^{(2)}_1]^2\}\\ &+ 2\mu\{[\kappa^{(1)}_2]^2[\kappa^{(2)}_2]^2 + [\kappa^{(1)}_1]^2[\kappa^{(2)}_1]^2\}. \end{aligned} \tag{E.3}$$

The right-hand side of (6.2) decays faster than the left-hand side. It can be shown that it has the asymptotic behavior

$$\Delta\mathbf{V}^{inc}(\xi) \to \frac{1}{\xi}, \qquad \xi \to \infty. \tag{E.4}$$

The asymptotic solution of (6.2) can be rewritten using (6.1) as

$$\Delta\mathbf{w}(\xi) = \xi^{-\nu}\mathbf{W}, \quad \nu = \frac{1}{2} \pm i\nu_0. \tag{E.5}$$

Then the Cauchy integral of the latter can be found in [20] to behave as

$$\frac{1}{\pi}\int_1^\infty \frac{\mathbf{W}d\zeta}{\zeta^\nu(\xi+\zeta)} \to \frac{\beta}{\xi^\nu}\mathbf{W}, \quad \xi \to \infty, \tag{E.6}$$

where $\beta = 1/\sin\nu\pi$. Therefore, using (E.1) and (E.4), as $\xi \to \infty$, the system (6.2) becomes

$$\frac{1}{\xi^\nu}(I_4 - \beta M_\infty)\mathbf{W} = \mathbf{0}. \tag{E.7}$$

The matrix in the above equation must have a zero determinant,

$$\det(I_4 - \beta M_\infty) = \begin{vmatrix} I_2 & -\beta m_\infty^- I_2 \\ -\beta m_\infty^+ I_2 & I_2 \end{vmatrix} = 1 - \beta^2 m_\infty^+ m_\infty^-. \tag{E.8}$$

This determines β, and hence, by its definition, the parameter ν_0:

$$\beta = \frac{1}{\sin\left(\frac{1}{2}\pm i\nu_0\right)\pi} = \frac{1}{\cosh(\pi\nu_0)} = \frac{1}{\sqrt{m_\infty^+ m_\infty^-}}. \tag{E.9}$$

Using one of the Dundurs parameters, which can be represented as

$$\beta_D = \frac{1 - [\kappa^{(2)}]^2 + \mu([\kappa^{(1)}]^2 + 1)}{\mu[\kappa^{(2)}]^2([\kappa^{(1)}]^2 - 1) + [\kappa^{(1)}]^2([\kappa^{(2)}]^2 - 1)} \tag{E.10}$$

(see [25]), as well as expressions (E.2) for $m_\infty^\pm$, we can write

$$\tanh(\pi\nu_0) = \beta_D. \tag{E.11}$$

Now it can be shown that there are two linearly independent vectors $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ such that $(I_4 - \beta M_\infty)\mathbf{W} = \mathbf{0}$. They are

$$\mathbf{W}^{(1)} = \begin{pmatrix} m \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{W}^{(2)} = \begin{pmatrix} 0 \\ m \\ 0 \\ 1 \end{pmatrix}, \quad m = \sqrt{\frac{m_\infty^-}{m_\infty^+}}. \tag{E.12}$$

It follows that as $\xi \to \infty$, the vector $\Delta\mathbf{w}$ behaves as

$$\Delta\mathbf{w}(\xi) \to \xi^{-1/2}\left[\left(\frac{A^+}{\xi^{i\nu_0}} + \frac{A^-}{\xi^{-i\nu_0}}\right)\mathbf{W}^{(1)} + \left(\frac{B^+}{\xi^{i\nu_0}} + \frac{B^-}{\xi^{-i\nu_0}}\right)\mathbf{W}^{(2)}\right], \qquad \xi \to \infty. \tag{E.13}$$

Rewriting the above equation in the form of (6.1) leads us to the following relationship:

$$\begin{aligned} \mathbf{D}^+ &= (A^+ + A^-)\mathbf{W}^{(1)} + (B^+ + B^-)\mathbf{W}^{(2)}, \\ \mathbf{D}^- &= -i[(A^+ - A^-)\mathbf{W}^{(1)} + (B^+ - B^-)\mathbf{W}^{(2)}]. \end{aligned} \tag{E.14}$$

Substituting (E.12) into (E.14) gives us a simple relationship between the components of $\mathbf{D}^\pm$,

$$D_1^\pm = \sqrt{\frac{m_\infty^-}{m_\infty^+}} D_3^\pm, \qquad D_2^\pm = \sqrt{\frac{m_\infty^-}{m_\infty^+}} D_4^\pm. \tag{E.15}$$

Acknowledgments. We are indebted to Profs. V. A. Borovikov and G. Mishuris for numerous enlightening discussions, and to an anonymous referee for insightful comments on the extinction theorems.

REFERENCES

[1] P. Bredif, C. Poideving, and O. Dupond, *A phased array technique for crack characterization*, in Proceedings of ECNDT 2006 (the European Conference on Nondestructive Testing); available online from http://www.ndt.net/article/ecndt2006/doc/Th.1.1.2.pdf.

[2] R. K. Chapman, J. Pearce, S. Burch, L. Fradkin, and M. Toft, *Recent in-house developments in theoretical modelling of ultrasonic inspection*, Insight, 49 (2007), pp. 93–97.

[3] L. A. Ahlberg and B. R. Tittmann, *Measurement techniques in elastic wave scattering experiments*, in Ultrasonics Symposium Proceedings, IEEE, New York, 1980, Vol. 2, pp. 842–846.

[4] B. R. Tittmann, *Scattering of elastic waves from simple defects in solids, A review*, Wave Motion, 5 (1983), pp. 299–306.

[5] K. M. Jaleel, N. N. Kishore, and V. Sundararajan, *Finite-element simulation of elastic-wave propagation in orthotropic composite-materials*, Materials Evaluation, 51 (1993), pp. 830–838.

[6] P. A. Lewis, J. A. G. Temple, E. J. Walker, and G. R. Wickham, *Calculation for diffraction coefficients for a semi-infinite crack embedded in an infinite anisotropic linearly elastic body*, Proc. R. Soc. Lond. A., 454 (1998), pp. 1781–1803.

[7] A. K. Gautesen, *Scattering by elastic quarter space*, Wave Motion, 7 (1985), pp. 557–568.

[8] A. K. Gautesen, *A geometrical theory of diffraction for crack-opening displacements*, Wave Motion, 10 (1988), pp. 393–404.

[9] A. K. Gautesen, *Scattering of a Rayleigh wave by an elastic wedge whose angle is greater than* 180 *degrees*, ASME J. Appl. Mech., 61 (2001), pp. 476–479.

[10] A. K. Gautesen, *On scattering of an SH-wave by a corner comprised of two different elastic materials*, Mech. Mat., 35 (2003), pp. 407–414.

[11] J. D. Achenbach, A. K. Gautesen, and H. McMaken, *Ray Methods for Waves in Elastic Solids: With Applications to Scattering by Cracks*, Pitman, New York, 1982.

[12] M. Lax, *Multiple scattering of waves.* II. *The effective field in dense systems,* Phys. Rev., 85 (1952), pp. 621–629.

[13] V. Kubzina, *Ultrasonic Phenomena with Application to Nondestructive Evaluation*, Ph.D. thesis, Department of Electrical, Electronic and Communications Engineering, London South Bank University, London, 2008.

[14] M. Comninou, *An overview of interface cracks*, Eng. Fracture Mech., 37 (1990), pp. 197–208.

[15] M. L. Williams, *The stresses around a fault or crack in dissimilar media*, Bull. Seismol. Soc. Amer., 49 (1959), pp. 199–404.

[16] F. Erdogan, *Stress distribution in a nonhomogeneous elastic plane with cracks*, J. Appl. Mech., 30 (1963), pp. 232–237.

[17] F. Erdogan, *Stress distribution in bonded dissimilar materials with cracks*, J. Appl. Mech., 32 (1965), pp. 403–410.

[18] J. R. Rice and G. C. Sih, *Plane problems of cracks in dissimilar media*, J. Appl. Mech., 32 (1965), pp. 418–423.

[19] Y. A. Antipov, *An exact solution of the* 3*D-problem of an interface semi-infinite plane crack*, J. Mech. Phys. Solids, 47 (1999), pp. 1051–1093.

[20] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[21] J. D. Achenbach, *Wave Propagation in Elastic Solids*, North–Holland, Amsterdam, 1973.

[22] A. Charalambopoulos, D. Gintides, and K. Kiriaki, *Radiation conditions for rough surfaces in linear elasticity*, Quart. J. Mech. Appl. Math., 55 (2002), pp. 421–441.

[23] J. A. DeSanto, *Exact boundary integral equations for scattering of scalar waves from perfectly reflecting infinite rough surfaces*, Wave Motion, 45 (2008), pp. 918–926.

[24] L. Cagniard, *Reflection and Refraction of Progressive Seismic Waves*, translated and revised by E. A. Flinn and C. H. Dix, McGraw–Hill, New York, 1962.

[25] J. Dundurs, *Effect of elastic constants on stress in a composite under plane deformation,* J. Compos. Mater., 1 (1967), pp. 310–322.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1334–1358

ON CENTER SUBSPACE BEHAVIOR IN THIN FILM EQUATIONS[*]

V. A. GALAKTIONOV[†] AND P. J. HARWIN[†]

Abstract. The large-time behavior of weak nonnegative solutions of the thin film equation (TFE) with absorption $u_t = -\nabla \cdot (|u|^n \nabla \Delta u) - |u|^{p-1}u$, with parameters $n \in (0,3)$ and $p > 1$, is studied. The standard free-boundary problem (FBP) with zero height, zero contact angle, and zero-flux conditions at the interface and bounded compactly supported initial data is considered. It is shown that there exists the *critical* absorption exponent $p_0 = 1 + n + \frac{4}{N}$ such that, for $p = p_0$, the asymptotic behavior of solutions $u(x,t)$ for $t \gg 1$ is represented by the well-known source-type solution of the pure TFE absorption, $u_s(x,t) = t^{-\beta N} F(y)$, $y = x/t^\beta$, with the exponent $\beta = \frac{1}{4+nN}$, which is perturbed by a couple of $\ln t$-factors. For $n = 1$, this behavior is associated with the center subspace for the rescaled linearized thin film operator and is given by $u(x,t) \sim (t \ln t)^{-\beta N} F(x/t^\beta (\ln t)^{-\beta N/4})$, with $\beta = \frac{1}{4+N}$, where $F(y) = \frac{1}{8(N+2)(N+4)}(a_*^2 - |y|^2)^2$ and the constant $a_* > 0$ depends on dimension N only. The $2m$th-order generalization of such TFEs with critical absorption is considered, and some local and asymptotic features of changing sign similarity solutions of the Cauchy problem are described. Our study is motivated by the phenomenon of logarithmically perturbed source-type behavior for the second-order porous medium equation (PME) with critical absorption $u_t = \nabla \cdot (u^n \nabla u) - u^p$ in $\mathbb{R}^N \times \mathbb{R}_+$, $p_0 = 1 + n + \frac{2}{N}$, $n \geq 0$, which has been known since the 1980s.

Key words. quasilinear thin film equation, critical absorption exponent, similarity solutions, asymptotic behavior

AMS subject classifications. 35K55, 35K65

DOI. 10.1137/060666275

1. Introduction: The model, motivation, and results.

Our goal is to describe some unusual asymptotic phenomena for *higher-order quasilinear* degenerate parabolic equations, in which the nonlinear interaction between operators involved deforms the scaling-invariant structure of solutions for large times. These delicate cases of asymptotic phenomena, such as *logarithmic perturbations* of fundamental or source-type solutions, have been known since the 1980s for quasilinear second-order reaction-diffusion equations. For *semilinear* higher-order parabolic equations, those phenomena can be detected by using spectral theory of non–self-adjoint operators and semigroup approaches. For *quasilinear* models, similar asymptotic patterns were unknown.

In the present paper, we introduce a new quasilinear parabolic model by adding to the standard thin film operator an extra absorption term. This creates a nonconservative evolution PDE, which enjoys a variety of logarithmically perturbed nonscaling asymptotics in both the free-boundary and the Cauchy problems. We then fix several similarities with simpler second-order diffusion-absorption models.

We begin with some physical motivation of such models.

1.1. On general thin film models: A class of conservative and nonconservative PDEs.

For a long time, modern thin film theory and application dealt with rather complicated nonlinear models. Typically, such models include the principal quasilinear fourth-order operator and several lower-order terms. For instance,

[*] Received by the editors July 28, 2006; accepted for publication (in revised form) November 6, 2008; published electronically February 19, 2009.
http://www.siam.org/journals/siap/69-5/66627.html

[†] Department of Math. Sci., University of Bath, Bath, BA2 7AY UK (vag@maths.bath.ac.uk, P.J.Harwin@maths.bath.ac.uk).

the *Benney equation* (1966) describes the nonlinear dynamics of the interface of two-dimensional liquid films flowing on a fixed inclined plane [2],

$$u_t + \tfrac{2\mathrm{Re}}{3}(u^3)_x + \varepsilon\big[\big(\tfrac{8\mathrm{Re}^2}{15}\,u^6 - \tfrac{2\mathrm{Re}}{3}\cot\theta\, u^3\big)u_x + \Sigma\, u^3 u_{xxx}\big]_x = 0, \tag{1.1}$$

where Re is the unit-order *Reynolds number* of the flow driven by gravity, σ is the rescaled *Weber number* (related to surface tension σ), θ is the angle of plane inclination to the horizontal, and $\varepsilon = \frac{d}{\lambda} \ll 1$, with d being the average thickness of the film and λ the wavelength of the characteristic interfacial disturbances. See [40].

Thin film equations (TFEs) can include nonpower nonlinearities. For instance, in multidimensional geometry, a typical example is

$$u_t + \nabla\cdot\big[\big(-G\,u^3 + \tfrac{BM\,u^2}{2P(1+B\,u)^2}\big)\nabla u\big] + S\,\nabla\cdot(u^3\nabla\Delta u) = 0, \tag{1.2}$$

which describes, in dimensionless form, the dynamics of a film in $\mathbb{R}^3$ subject to the actions of thermocapillary, capillary, and gravity forces. Here, G, M, P, B, and S are the gravity, Marangoni, Prandtl, Biot, and inverse capillary numbers, respectively. For more on Marangoni instability in such TFE models, see [38].

The above conservative PDEs preserve the finite mass of thin films. Nonconservative TFEs occur for evaporating/condensing films and via other effects [39, 28]. Actually, the first study of the vapor thrust effects in the Rayleigh–Taylor instability of an evaporating liquid-vapor interface above a hot horizontal wall was performed by Bankoff in 1961. His stability analysis in 1971 of an evaporating thin liquid film on a hot inclined wall extended earlier results of Yih (1955, 1963) and Benjamin (1957). The history and detailed derivation of models of (a) an evaporating thin film and (b) a condensing thin film can be found in [39, pp. 946–949]. A typical TFE of that type in one dimension is as follows [39, p. 949]:

$$u_t + \tfrac{\bar{E}}{u+K} + \tfrac{1}{3}\tfrac{1}{C}\big(u^3 u_{xxx}\big)_x + \big\{\big[\tfrac{A}{u} + \tfrac{\bar{E}^2}{D}\big(\tfrac{u}{u+K}\big)^3 + \tfrac{KM}{\mathrm{Pr}}\big(\tfrac{u}{u+K}\big)^2\big]u_x\big\}_x = 0. \tag{1.3}$$

Here, the six terms represent, respectively, the rate of volumetric accumulation, the mass loss, the stabilization capillary, van der Waals, vapor thrust, and thermocapillary effects. In the second absorption-like term, $\bar{E}$ is the scaled evaporation number and K is the scaled interfacial thermal resistance that physically represents a temperature jump from the liquid surface temperature to the uniform temperature of the saturated vapor. D is a unit-order scaled ratio between the vapor and liquid densities.

Another origin of nonconservative TFEs with more complicated nondivergent operators is the study of flows on a rotating disc (centrifugal spinning as an efficient mean of coating planar solids with thin films). This gives extra absorption-like, spatially nonautonomous terms in the equations written in radial geometry, e.g., [39, p. 955]

$$\begin{aligned} &u_t + \tfrac{2}{3}E + \tfrac{1}{3r}\big[r^2u^3 + \varepsilon\mathrm{Re}\big(\tfrac{5}{12}Er^2u^4 - \tfrac{34}{105}r^2u^7\big)\big] \\ &+ \tfrac{\varepsilon}{3}\big\{\mathrm{Re}\big(\tfrac{2}{5}r^3u^6 - r\tfrac{1}{F^2}u^3\big)u_r + r\tfrac{1}{C}u^3\big[\tfrac{1}{r}(ru_r)_r\big]_r\big\}_r = 0. \end{aligned} \tag{1.4}$$

Here E is again the evaporation number, F is the Froude number, and $\varepsilon = \frac{h_0}{L}$ is a small parameter. Observe a rather complicated combination of various absorption- and reaction-like nondivergent terms (with different nonlinear powers u^3, u^4, and u^7) in the first line of (1.4). Various exact solutions of nonconservative TFEs can be found in [25, Ch. 3], where more references and a survey on TFE theory are given.

Modern nonlinear parabolic theory and application to thin film models demand better understanding of the interaction of various nonlinear terms and operators of different orders that can create rather complicated spatiotemporal patterns and dissipative structures. We chose one particular but special case of center subspace behavior that will be shown to have rather robust mathematical significance.

1.2. Basic limit model: The TFE with absorption. We study the large-time asymptotic behavior of nonnegative solutions of the *thin film equation* (TFE) *with absorption* (for convenience, we also examine solutions of changing sign)

$$u_t = -\nabla \cdot (|u|^n \nabla \Delta u) - |u|^{p-1} u, \tag{1.5}$$

where $n > 0$ and $p > 1$ are fixed exponents. Here we use the simplest second term which is not a differential operator but is represented by just a power function. Our main goal is to justify that in the critical case

$$p_0 = 1 + n + \tfrac{4}{N} \tag{1.6}$$

various solutions of (1.5) exhibit a complicated asymptotic behavior with some logarithmic corrections $\ln t$ for $t \gg 1$.

We have chosen the nonconservative equation (1.5) for simplicity and for better presentation of our mathematical tools. We claim that similar phenomena are quite general and appear also in various conservative models. Actually, the logarithmic correction $\sim (\ln \frac{1}{t})^{-1/7}$ in the behavior for large enough t was rigorously observed [27] for the relaxed conservative thin film model consisting of two operators,

$$u_t + (u^3 u_{xxx})_x + (u^n u_{xxx})_x = 0, \quad \text{with} \quad 0 < n < 3 \quad (u \geq 0), \tag{1.7}$$

where the first term with u^3 corresponds to Reynolds's equation from lubrication theory. It was shown that, for concentrated enough initial data, in a certain intermediate time-range, the propagation rate is as follows:

$$\text{meas}\,\{u(x,t) > 1\} \sim \big(\tfrac{t}{\ln \frac{1}{t}}\big)^{\frac{1}{7}}, \tag{1.8}$$

where the usual scaling-invariant factor $t^{\frac{1}{7}}$ is associated with a standard dimensional analysis. Here, the log-correction is a result of a delicate interaction of two scaling invariant operators in (1.7). We believe that (1.8), proved in [27] rigorously, can be put into a framework of a center manifold calculus (though a justification can be extremely hard).

Log-corrections were observed for the limit stable Cahn–Hilliard equation [20, section 5.4]

$$u_t = -\Delta^2 u + \Delta(|u|^{p-1} u), \quad \text{with} \quad p = 1 + \tfrac{2}{N}. \tag{1.9}$$

For the semilinear case $n = 0$ in the TFE (1.5), such logarithmically perturbed asymptotic are also well known and admit a rigorous mathematical treatment [22].

Thus, we consider for (1.5) the standard free-boundary problem (FBP) with *zero height*, *zero contact angle*, and *zero-flux* (conservation of mass) conditions

$$u = \nabla u = \nu \cdot (u^n \nabla \Delta u) = 0 \tag{1.10}$$

at the singularity surface (interface) $\Gamma_0[u]$, which is the lateral boundary of $\operatorname{supp} u$ with the outward unit normal ν. Bounded, smooth, and compactly supported initial data

$$u(x,0) = u_0(x) \quad \text{in } \Gamma_0[u] \cap \{t=0\} \tag{1.11}$$

are added to complete a suitable functional setting of the FBP. As usual, we assume that these three free-boundary conditions give a correctly specified problem for the fourth-order parabolic equation, at least for sufficiently smooth and bell-shaped initial data, e.g., in the radial setting.

Returning to basics of thin film theory, earlier references on derivation of the pure fourth-order TFE

$$u_t = -\nabla \cdot (|u|^n \nabla \Delta u) \tag{1.12}$$

and related models can be found in [29, 42], where the first analysis of some self-similar solutions for $n=1$ was performed. Source-type similarity solutions of (1.12) for arbitrary n were studied in [7] for $N=1$ and in [21] for the equation in $\mathbb{R}^N$. More information on similarity and other solutions can be found in [5, 4, 11]. In general, the TFEs are known to admit nonnegative solutions constructed by special "singular" parabolic approximations of the degenerate nonlinear coefficients; see the pioneering paper [3], various extensions in [30, 15, 16, 34, 45], and the references therein. In what follows we study the asymptotic behavior of sufficiently "strong" weak solutions of the TFEs, which satisfy necessary regularity and other assumptions; see also the survey paper [1]. Notice that regularity theory for the TFEs is not fully developed, especially in the nonradial N-dimensional geometry and for solutions of changing sign, so we will need to impose extra formal requirements, which are necessary for justifying our asymptotic approaches.

Let us mention other well-established and related conservative thin film models with extra lower-order terms describing the dynamics of thin films of viscous fluids in the presence of two competing forces; see [9]. For $N=1$, typical *quasilinear* TFEs are

$$u_t = -(uu_{xxx} + u^3 u_x)_x \quad (u \geq 0), \tag{1.13}$$

and the general equation with power nonlinearities is

$$u_t = -(u^n u_{xxx})_x - (u^m u_x)_x \quad (u \geq 0). \tag{1.14}$$

We refer the reader to papers [17, 18] and the book [25, Ch. 3] as sources of a large number of further references and results of TFE theory and application.

In addition, our extra motivation of the TFE model like (1.5) is mathematical and is associated with the previous investigations of the quasilinear diffusion-absorption PDEs.

1.3. A mathematical motivation: The PME with critical absorption. Second-order quasilinear parabolic equations with absorption are well known in combustion theory. A key model is the *porous medium equation* (PME) *with absorption*

$$u_t = \nabla \cdot (u^n \nabla u) - u^p \quad \text{in } \mathbb{R}^N \times \mathbb{R}_+ \quad (u \geq 0), \tag{1.15}$$

where $n > 0$ and p are fixed exponents. A special interest to such equations was motivated by localized similarity solutions introduced by L.K. Martinson and K.B. Pavlov at the beginning of the 1970s. Mathematical theory of such PDEs was developed by A.S. Kalashnikov a few years later; see his survey [31] for the full history. Besides new phenomena of localization and interface propagation, for more than twenty years, the PME with absorption (1.15) became a crucial model for determining various asymptotic patterns, which can occur for large times $t \gg 1$ or close to finite-time extinction as $t \to T^-$ (for $p < 1$). For (1.15), there are a few parameter ranges with different asymptotics:

$$p > p_0 = 1 + n + \tfrac{2}{N}, \quad p = p_0, \quad 1 + n < p < p_0, \quad p = 1 + n,$$
$$1 < p < 1 + n, \quad p = 1, \quad 1 - n < p < 1, \quad p = 1 - n, \quad p < 1 - n,$$

etc.; see references and details in [26, Chs. 5, 6].

The most interesting and unusual transitional behavior for (1.15) occurs at the first *critical* (or *Fujita*) absorption exponent

$$p_0 = 1 + n + \tfrac{2}{N}. \tag{1.16}$$

In this case (see details and references in [26, p. 83]), the asymptotic behavior as $t \to \infty$ of nonnegative compactly supported solutions of (1.15) is described by the logarithmically perturbed source-type solution of the pure PME,

$$u(x,t) = (t \ln t)^{-\beta N}[F(x/t^{\beta}(\ln t)^{-\beta n/2}) + o(1)], \quad \text{where } \beta = \tfrac{1}{2+nN}. \tag{1.17}$$

Without the logarithmic factors and the $o(1)$-term, the right-hand side is indeed the famous *Zel'dovich–Kompaneetz–Barenblatt* (ZKB) similarity source-type solution of the pure PME $u_t = \nabla \cdot (u^n \nabla u)$, which has the form

$$u_s(x,t) = t^{-\beta N} F(y), \;\; y = x/t^{\beta}, \quad \text{with } F(y) = \left[\tfrac{n\beta}{2}(a^2 - |y|^2)_+\right]^{\frac{1}{n}}, \tag{1.18}$$

where $a > 0$ is an arbitrary scaling parameter. This explicit solution dates back to the 1950s. In the class of solutions of changing sign, (1.15) admits a countable sequence of critical exponents, where the patterns contain similar logarithmic time-factors [23].

1.4. Outline of the paper: Logarithmically perturbed patterns for the TFE with absorption. In section 2 we show that similar logarithmically perturbed source-type patterns exist for the TFE with absorption (1.5), with the critical exponent (1.6). In this case, the source-type solutions of the TFE (1.12) take the form

$$u_s(x,t) = t^{-\beta N} F(y), \quad y = x/t^{\beta}, \quad \text{with } \beta = \tfrac{1}{4+nN}, \tag{1.19}$$

where $F(y) \geq 0$ is a radially symmetric compactly supported solution of the PDE [7, 21]

$$\mathbf{A}(F) \equiv -\nabla \cdot (F^n \nabla \Delta F) + \beta \nabla F \cdot y + \beta N F = 0. \tag{1.20}$$

In the case $n = 1$, the similarity profile for the FBP is given explicitly,

$$F(y) = c_0(a^2 - |y|^2)^2, \quad c_0 = \tfrac{1}{8(N+2)(N+4)}, \quad a > 0, \tag{1.21}$$

and was first constructed in [42]. Figure 1 shows profiles $F(y)$ for $N = 1$ in four cases $n = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and 1. The profiles are normalized by their values at $y = 0$, so $F(0) = 1$.

First, for $n = 1$, relying on the explicit representation (1.21) and good spectral properties of the corresponding self-adjoint linearized rescaled operator, we show that,

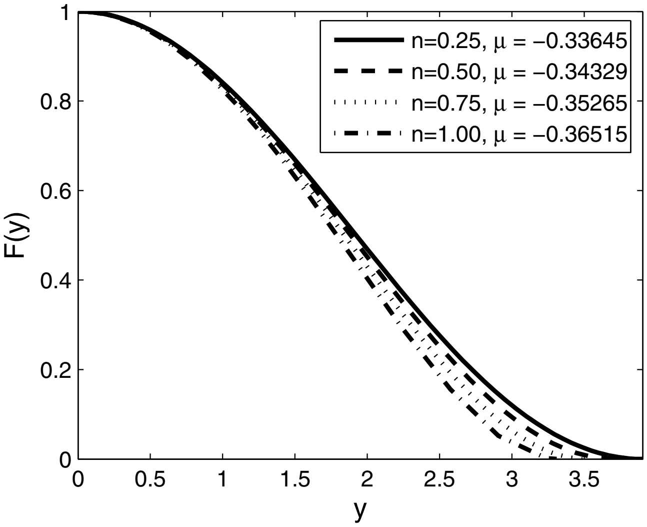


FIG. 1. *The similarity profiles $F(y)$ of* (1.20) *for $N = 1$ in four cases $n = \frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, and* 1*; $F(0) = 1$, $\mu = F''(0)$.*

for $p = p_0 = 2 + \frac{4}{N}$, the TFE with absorption (1.5) admits asymptotic patterns of the following form:

$$u(x,t) \sim (t \ln t)^{-\beta N} F_*(x/t^{\beta}(\ln t)^{-\beta N/4}) \quad \left(\beta = \tfrac{1}{4+N}\right). \tag{1.22}$$

Here F_* is a fixed rescaled profile from the family (1.21) with a uniquely chosen parameter $a = a_* > 0$ that depends on N only. We also present evidence that similar logarithmic factors can occur for arbitrary $n > 0$, but this does not lead to self-adjoint linearized operators and explicit mathematics. On the other hand, for the semilinear case $n = 0$, i.e., for the fourth-order parabolic equation written for solutions of changing sign

$$u_t = -\Delta^2 u - |u|^{p-1} u, \tag{1.23}$$

the critical behavior like (1.22) is known to occur at the critical exponent $p = 1 + \frac{4}{N}$ [22], which is precisely (1.6) with $n = 0$. In this case, the center manifold analysis also uses spectral properties of a non–self-adjoint linear operator studied in [13, section 2].

In section 3 we briefly describe the essence of the easier supercritical case $p > p_0$. Very singular similarity solutions (VSSs) in the subcritical one $p \in (n+1, p_0)$ will be studied in a forthcoming paper.

In section 4, we explain how the critical asymptotic behavior occurs for the $2m$th-order TFE with absorption

$$u_t = (-1)^{m+1} \nabla \cdot (u^n \nabla \Delta^{m-1} u) - u^p, \quad m \geq 2, \tag{1.24}$$

where the critical absorption exponent is

$$p_0 = 1 + n + \tfrac{2m}{N}, \tag{1.25}$$

and again $n = 1$ leads to a simpler self-adjoint case.

In section 5 we discuss similar local and global asymptotics for the Cauchy problem admitting *maximal regularity* solutions of changing sign.

Finally, let us note that (1.6) (and (1.25) for equation (1.24)) is the critical *Fujita* exponent of the TFE with source

$$u_t = -\nabla \cdot (|u|^n \nabla \Delta u) + |u|^p \quad (n > 0,\ p > 1);$$

i.e., for $p \in (n+1, p_0]$, all solutions with arbitrarily small initial data $u_0(x)$, where $\int u_0 > 0$, blow up in finite time [24].

2. Rescaled equation and center subspace behavior.

2.1. To the style of the analysis. For convenience of the reader, we must emphasize from the beginning that all our final conclusions on center subspace behaviors detected below are *mathematically formal* when we deal with the quasilinear case $n > 0$. The semilinear case $n = 0$ is easier and admits a rigorous treatment by invariant manifold theory [22]. It is then worth mentioning that there is no hope that such asymptotics can admit a reasonably simple rigorous treatment. We recall that even for the second-order model (1.15) with $n > 0$, there is no full center manifold justification of the main results that were proved by essential use of the maximum principle and comparison-barrier techniques; see [26, Ch. 4]. Some of the asymptotic patterns for (1.15) of center subspace type turned out to be very complicated [23]. As we will show, the main difficulty is not a proper spectral theory of linearized operators (this is justified in many cases) but a justification of the center subspace behavior associated with such singular operators. On the other hand, we always clearly indicate the rigorous steps and split the whole approach into a sequence of standard steps. We would be very pleased if some of our formal results and discussions would attract the attention of experts in these areas of differential equations.

Thus, in what follows, we use by implication the following rule:

(i) all conclusions concerning spectral and other properties of self-adjoint singular elliptic and ordinary differential operators are *rigorous* (or can be made rigorous after sometimes technical manipulations; for non–self-adjoint cases we are not certain, and extra analysis is necessary); and

(ii) further extensions via the above spectral properties to describe the behavior for TFEs close to center subspaces and various matching procedures are mathematically *formal*.

2.2. Rescaled equation. We begin with rescaling the PDE (1.5) with the critical exponent (1.6) according to the time-factors of the source-type solution (1.19), i.e., by setting

$$u(x,t) = (1+t)^{-\beta N} v(y,\tau), \quad y = x/(1+t)^{\beta}, \quad \tau = \ln(1+t), \tag{2.1}$$

which leads to the following autonomous rescaled equation in $\mathbb{R}^N \times \mathbb{R}_+$:

$$v_\tau = \mathbf{A}(v) - v^p, \tag{2.2}$$

where $\mathbf{A}$ is the operator specified in (1.20). We first need to check that a simple stabilization as $\tau \to +\infty$ to a nontrivial stationary solution in (2.2) is not possible.

Proposition 2.1. *The stationary equation*

$$\mathbf{A}(g) - g^p = 0 \tag{2.3}$$

does not have a nontrivial compactly supported nonnegative solution of the FBP.

Proof. Indeed, integrating (2.3) over $\operatorname{supp} g$ yields $\int g^p(y)\,\mathrm{d}y = 0$. □

This means that the only bounded nonnegative equilibrium for the dynamical system (2.2) is trivial:

$$g(y) \equiv 0 \quad \text{in } \mathbb{R}^N. \tag{2.4}$$

In order to detect the actual nonstationary asymptotic behavior, we next perform a second rescaling by introducing the as yet unknown positive function $b(\tau)$,

$$v(y,\tau) = b(\tau)w(\zeta,\tau), \quad \zeta = y/b^{\frac{n}{4}}(\tau), \tag{2.5}$$

to get the following perturbed equation:

$$w_\tau = \mathbf{A}(w) + \tfrac{b'}{b}\,\mathbf{C}w - b^{p-1}w^p, \quad \text{where } \mathbf{C}w \equiv \tfrac{n}{4}\,\nabla w \cdot \zeta - w. \tag{2.6}$$

2.3. Linearization. Roughly speaking, in order to detect the asymptotic behavior according to (2.5), we can use the estimate

$$b(\tau) \approx \|v(\cdot,\tau)\|_\infty \to 0 \quad \text{as } \tau \to \infty, \tag{2.7}$$

so that $\|w(\cdot,\tau)\|_\infty \approx 1$ for $\tau \gg 1$. On the other hand, in the radial setting, it is convenient to use $b(\tau)$ for the scaling of the support of the solution $w(\zeta,\tau)$ to get that it approaches the unit ball B_1 as $\tau \to \infty$; see below.

We next perform the linearization by setting

$$w(\zeta,\tau) = F(\zeta) + Y(\zeta,\tau), \tag{2.8}$$

where F is a rescaled similarity profile from the family (1.21). Then Y solves the following rescaled equation:

$$Y_\tau = \mathbf{A}'(F)Y + \tfrac{b'}{b}\,\mathbf{C}F - b^{p-1}F^p + \mathbf{D}(Y) - b^{p-1}[(F+Y)^p - F^p], \tag{2.9}$$

where $\mathbf{A}'(F)Y$ is the formal Fréchet derivative of $\mathbf{A}$ at F,

$$\mathbf{A}'(F)Y = -\nabla \cdot (F^n \nabla \Delta Y) - \nabla \cdot (nF^{n-1}Y\nabla\Delta F) + \beta\nabla Y \cdot y + \beta N Y, \tag{2.10}$$

and $\mathbf{D}(Y)$ is a higher-order perturbation, which is quadratic in $Y \to 0$ on smooth functions. Using the elliptic equation (1.20) for F, on integration,

$$F^n\nabla\Delta F = \beta F y \quad \Longrightarrow \quad \mathbf{A}'(F)Y = -\nabla \cdot (F^n\nabla\Delta Y) + (1-n)\beta\nabla \cdot (Y\zeta). \tag{2.11}$$

2.4. The self-adjoint case $n = 1$. It follows from (2.11) that $n = 1$ is a special case, where the last term vanishes. We fix $a = 1$ in (1.21), so that the linearized operator is

$$\mathbf{A}'(F)Y = -\nabla \cdot (F\nabla\Delta Y) \equiv -c_0\nabla \cdot ((1-|\zeta|^2)\nabla\Delta Y), \quad y \in B_1 = \{|\zeta| < 1\}. \tag{2.12}$$

One can see that it can be written in the form

$$\mathbf{A}'(F)Y = -c_0\,\tfrac{1}{\rho(|\zeta|)}\,[\Delta(a(|\zeta|)\Delta Y) + 2N\Delta Y], \quad \text{where } a(|\zeta|) = 1 - |\zeta|^2 = \tfrac{1}{\rho(|\zeta|)}, \tag{2.13}$$

so, in the topology of $L^2_\rho(B_1)$, operator (2.12) is symmetric in $C_0^\infty(B_1)$ with good coefficients and hence admits self-adjoint extensions. Next, using classical theory

[10], we specify properties of its unique Friedrichs self-adjoint extension. Its domain is constructed by completing $C_0^\infty(B_1)$ in the norm induced by its positive quadratic form (corresponding to the operator $-\mathbf{A}'(F) - c_0\Delta > 0$)

$$\langle Y, W\rangle_* \equiv c_0 \textstyle\int_{B_1} [a\Delta Y \Delta W - (2N-1)\nabla Y \cdot \nabla W].$$

The intersection of this Hilbert space with the domain of the maximal adjoint operator $D((\mathbf{A}'(F))^*) = \{v \in L^2_\rho : \ \mathbf{A}'(F)v \in L^2_\rho\}$ defines the domain of the self-adjoint extension, which we denote by $D(\mathbf{A}'(F)) = H^4_{\rho,0}$. In particular, for any $v \in H^4_{\rho,0}$, there holds

$$v = 0 \quad \text{on } \partial B_1, \quad \text{and} \quad \textstyle\int_{B_1} a(\Delta v)^2 < \infty,$$

so that $H^4_{\rho,0} \subset H^2_{\rho,0}$. Consider the corresponding eigenvalue problem written in the form

$$-c_0[\Delta(a(|\zeta|)\Delta\psi) + 2N\Delta\psi] = \rho\lambda\psi \quad \text{in } H^4_{\rho,0}. \tag{2.14}$$

Since the embeddings of the corresponding functional spaces $H^2_{a,0}$ and H^1_0 into L^2_ρ are compact [36, p. 63], we have that the spectrum $\sigma(\mathbf{A}'(F))$ is real and discrete.

For our purposes, it suffices to detect the eigenvalues and eigenfunctions in the radial (ODE) setting with the single spatial variable $r = |\zeta| > 0$. The extension to the elliptic setting is performed by using the polar coordinates $\zeta = (r, \sigma)$ in B_1,

$$\Delta = \Delta_r + \tfrac{1}{r^2}\,\Delta_\sigma, \tag{2.15}$$

where Δ_σ is the Laplace–Beltrami operator on the unit sphere $S^{N-1} = \partial B_1$ in $\mathbb{R}^N$. Δ_σ is a regular operator with a discrete spectrum in $L^2(S^{N-1})$ (each eigenvalue repeated as many times as its multiplicity),

$$\sigma(-\Delta_\sigma) = \{\nu_k = k(k+N-2), \ k \geq 0\}, \tag{2.16}$$

and an orthonormal, complete, closed subset $\{V_k(\sigma)\}$ of eigenfunctions, which are homogeneous harmonic kth order polynomials restricted to S^{N-1}. We plug (2.15) into (2.13), where all the coefficients are radial functions, and use the separation of variables

$$\psi(r, \sigma) = R(r)V_k(\sigma) \tag{2.17}$$

to solve the eigenvalue problem (2.14). For each fixed ν_k, we then arrive at a radial eigenvalue problem for R, which is similar to that discussed below.

Thus we take $k = 0$ in (2.17) and consider the radially symmetric eigenvalue problem (2.14). For $N = 1$, this problem was studied in [8], where further references are given. It is not difficult to check that the radial operator $\mathbf{A}'(F)$ has the discrete spectrum

$$\sigma(\mathbf{A}'(F)) = \{\lambda_k = c_0 k(k+2)(k+N)(k+N+2), \ k = 0, 2, 4, \ldots\}, \tag{2.18}$$

where each eigenfunction ψ is a $(k{+}2)$th-order polynomial,

$$\psi_k(r) = b_k(r^{k+2} + \cdots + d_k) \quad (\psi_k(1) = 0), \tag{2.19}$$

where $\{b_k\}$ are normalization constants, so that the eigenfunction subset $\{\psi_k\}$ is orthonormal in L^2_ρ. In particular,

$$\psi_0(r) = b_0(r^2-1) > 0, \quad b_0 = -\sqrt{\tfrac{N+2}{2\omega_N}} \quad (\lambda_0 = 0), \tag{2.20}$$

where $\omega_N = \frac{2\pi^{N/2}}{N\Gamma(N/2)}$ is the volume of the unit ball in $\mathbb{R}^N$. Such polynomials are complete and closed in typical weighted L^p-spaces (a standard functional analysis result; see [13, section 2.3] for details), and this justifies the equality in (2.18). Moreover, we then can use the eigenfunction expansion with the orthonormal eigenfunction subset $\{\psi_k\}$ to deal with solutions of the corresponding PDE.

We next consider the rescaled equation (2.9), which for $n = 1$ takes the form

$$Y_\tau = \mathbf{A}'(F)Y + \tfrac{b'}{b}\,\mathbf{C}F - b^{p-1}F^p - \nabla\cdot(Y\nabla\Delta Y) - b^{p-1}[(F+Y)^p - F^p]. \tag{2.21}$$

We deal with strong radially symmetric solutions of (2.21), where we now choose the normalization function $b(\tau)$ in (2.5) such that

$$\operatorname{supp} w(\cdot,\tau) = B_1 \quad \text{for } \tau \gg 1. \tag{2.22}$$

According to (2.21), we then need to assume that $b(\tau)$ is smooth, at least for large τ, though this requirement can be weaken by using a weak (integral) form of the PDE. We now use the converging (in L^2_ρ and in the corresponding Sobolev class) eigenfunction expansion of the radial solution

$$Y(\zeta,\tau) = \textstyle\sum_{k\ge 0} a_k(\tau)\psi_k(\zeta) \tag{2.23}$$

to study the corresponding center subspace behavior for the nonlinear operator $\mathbf{A}$. This part of our asymptotic analysis is formal.

Thus substituting (2.23) into (2.21) and projecting onto ψ_0 in L^2_ρ, we have that the first coefficient satisfies the following perturbed ODE:

$$a_0' = -\gamma_1\tfrac{b'}{b} - \gamma_2 b^{p-1} + \dots, \ \text{ where } \ \gamma_1 = -\langle \mathbf{C}F, \psi_0\rangle_\rho > 0, \ \gamma_2 = \langle F^p, \psi_0\rangle_\rho > 0. \tag{2.24}$$

We omit in (2.24) the higher-order terms, assuming that, for this type of behavior, the nonautonomous perturbations are the leading ones. The signs of the coefficients $\gamma_{1,2}$ in (2.24) are essential and are easily checked by integration.

It follows from (2.4) and (2.22) that $b(\tau) \to 0$ as $\tau \to \infty$, so

$$\tfrac{b'(\tau)}{b(\tau)} \quad \text{is not integrable at } \tau = \infty.$$

Therefore, in order to have a uniformly bounded expansion coefficient $a_0(\tau)$, we need to suppose that two terms on the right-hand side of (2.24) annul each other asymptotically, so that, up to an integrable perturbation,

$$\tfrac{b'}{b} = -\tfrac{\gamma_2}{\gamma_1}\,b^{p-1} + \dots \quad \text{for } \tau \gg 1. \tag{2.25}$$

This gives the following necessary condition for the existence of such behavior:

$$b(\tau) = \gamma_*\tau^{-\frac{1}{p-1}} + \dots, \quad \text{where } \gamma_* = \left[\tfrac{(p-1)\gamma_2}{\gamma_1}\right]^{-\frac{1}{p-1}}. \tag{2.26}$$

Returning to the original variables $\{x,t,u\}$, from (2.26) we obtain the asymptotic pattern (1.22). The rescaled profile F_* is uniquely determined from (1.21) with $a_* = \gamma_*^{n/4}$.

2.5. Arbitrary $n \in (0, \frac{3}{2})$. This non–self-adjoint case is more difficult. Consider the linearized operator (2.11) for $n \neq 1$, where $F > 0$ is the radial solution of the ODE (1.20) in B_1; see [21] for existence, uniqueness, and asymptotics. Then, for $n < \frac{3}{2}$ [21],

$$F(\zeta) \sim (1 - |\zeta|)^2 \quad \text{as } |\zeta| \to 1^-. \tag{2.27}$$

Notice that there exists a one-parameter family of the solutions given by

$$F_a(\zeta) = a^{\frac{4}{n}} F(\tfrac{\zeta}{a}), \quad a > 0. \tag{2.28}$$

First, we claim that, for $n \neq 1$, operator (2.11) is not symmetric in L^2_ρ for any positive weight ρ in B_1; see the appendix. Second, we have that

$$\psi_0(\zeta) = \tfrac{\mathrm{d}}{\mathrm{d}a} F_a(\zeta)|_{a=1} \equiv \tfrac{4}{n} F - \nabla F \cdot \zeta \tag{2.29}$$

is a positive eigenfunction of (2.11) corresponding to $\lambda_0 = 0$. Observe that, with respect to the regularity, this eigenfunction well corresponds to that for $n = 1$; cf. (2.20). Moreover, it follows that, close to the singular point $|\zeta| = 1$, the radial part of (2.11) is governed by the singular (at ∂B_1) higher-order operator

$$L_4 Y = -(s^{2n} Y''')', \quad s = 1 - |\zeta|, \tag{2.30}$$

which is symmetric in a weighted H^{-1} topology (but we need a result in L^2). Solving the problem $L_4 Y = g$ with natural conditions at the point $s = 1$, which is assumed to be regular, we obtain, up to compact perturbations, that

$$L_2 Y \equiv -Y'' \sim \int^s s^{-2n} \int^s g \equiv L_* g \quad \Longrightarrow \quad Y \sim L_2^{-1} L_* g, \tag{2.31}$$

where L_2^{-1} is a compact operator in L^2. It is easy to check that the integral operator L_* is bounded in L^2 for

$$n < \tfrac{3}{4}, \tag{2.32}$$

and then $L_2^{-1} L_*$ is compact in L^2 as the product of a compact and a bounded operator. Therefore, $\mathbf{A}'(F)$ has discrete spectrum in the parameter range (2.32). This is not an optimal result since, as we have seen, the discreteness of the spectrum remains valid for $n = 1$. We use this analysis as a simple illustration of the fact that the spectrum is usually discrete in the nonsymmetric case.

Thus $0 \in \sigma(\mathbf{A}'(F))$ is an isolated eigenvalue. There is numerical evidence that the spectrum is discrete for all $n \in (0, \frac{3}{2})$; see [8], where, moreover, the first six eigenvalues turned out to be *real* for $N = 1$. Possibly this might mean that in a special topology of sequences as l^2 (not related to any of L^2_ρ) the linearized operator can be treated as symmetric and self-adjoint; cf. an example in [13]. For $n = 0$ in any dimension $N \geq 1$, the whole spectrum is proved to be real. We refer the reader to [13, section 2], where this and other $2m$th-order operators were studied in $L^2_\rho(\mathbb{R}^N)$, i.e., for the Cauchy (not a free-boundary) problem.

The rest of our study is formal. Once in the radial setting there exists the center subspace of $\mathbf{A}'(F)$; we are looking for a (formal) center subspace patterns for (2.9)

$$Y(\zeta, \tau) = a_0(\tau) \psi_0(\zeta) + \dots . \tag{2.33}$$

We assume center subspace dominance in the behavior, so, as usual, other terms in this expansion are assumed to be negligible for $\tau \gg 1$. Substituting (2.33) into (2.9), we next find the projection onto the corresponding adjoint eigenfunction ψ_0^*. In general, such an analysis becomes rigorous if we establish existence of complete, closed, and biorthonormal eigenfunction subsets $\{\psi_k\}$ and $\{\psi_k^*\}$. This is an open problem except for the case $n = 1$ above and $n = 0$ studied in [13]. We do not deal with the adjoint operator $\mathbf{A}'^*(F)$ in this formal asymptotic analysis. The projection onto ψ_0^* yields the perturbed ODE (2.24), where the same coefficients $\gamma_{1,2}$ are determined via the standard dual L^2 product, where ψ_0 is replaced by ψ_0^*. This formally leads to the same asymptotics (2.26).

The range $n \in [\frac{3}{2}, 3)$. The center subspace analysis applies also for larger n's. The asymptotics of similarity profiles change at $n = \frac{3}{2}$, where, instead of (2.27),

$$F(\zeta) \sim (1-|\zeta|)^2 \big[\tfrac{3}{4}\,\beta|\ln(1-|\zeta|)|\big]^{\frac{2}{3}} \quad \text{as } |\zeta| \to 1; \tag{2.34}$$

see [21]. On the other hand, for $n \in (\frac{3}{2}, 3)$,

$$F(\zeta) \sim (1-|\zeta|)^{\frac{3}{n}} \quad \text{as } |\zeta| \to 1. \tag{2.35}$$

This regularity is sufficient for determining the corresponding eigenfunction and the logarithmic behavior.

For $n \geq 3$, the zero contact angle FBP does not provide us with a proper interesting evolution; see [21].

3. On the supercritical parameter range $p > p_0$.

3.1. Exponentially perturbed dynamical system for $p > p_0$. Let us explain what we expect for $p > p_0$ in (1.5). In terms of the rescaled function

$$u(x,t) = (1+t)^{-\frac{N}{4+nN}}\, v(y,\tau), \quad \tau = \ln(1+t), \tag{3.1}$$

the equation takes the form

$$v_\tau = -\nabla \cdot (v^n \nabla \Delta v) + \tfrac{1}{4+nN}\, y \cdot \nabla v + \tfrac{N}{4+nN}\, v - \mathrm{e}^{-\gamma\tau} v^p, \tag{3.2}$$

where $\gamma = \frac{N(p-p_0)}{4+nN} > 0$ if $p > p_0$. Therefore, the absorption term $-u^p$ in (1.5) generates an exponentially small perturbation in the rescaled equation (3.2). Hence one can expect the convergence as $\tau \to \infty$ to the rescaled similarity profile F in (1.19) of the limit mass, though the passage to the limit in (3.2) generates a number of technical difficulties. Here (3.2) is formally an exponentially small perturbation of the autonomous rescaled TFE

$$v_\tau = \mathbf{A}(v) \equiv -\nabla \cdot (v^n \nabla \Delta v) + \tfrac{1}{4+nN}\, y \cdot \nabla v + \tfrac{N}{4+nN}\, v. \tag{3.3}$$

As usual, we gain an extra advantage in the case $n = 1$.

3.2. The gradient case $n = 1$. It is known that, for $n = N = 1$, the rescaled TFE (3.3) is a gradient system [12]. Let us construct an "approximate" Lyapunov function for strong solutions of the FBP in $\mathbb{R}^N$. Namely, we write down (3.2) in the form

$$v_\tau = \nabla \cdot \big[v \nabla \big(-\Delta v + \tfrac{1}{2(4+N)}\, |y|^2\big)\big] + \mathrm{e}^{-\gamma\tau} v^p \tag{3.4}$$

and multiply in $L^2(\mathbb{R}^N)$ by $(-\Delta_v)^{-1} v_\tau$, where, by definition,

$$(-\Delta_v)^{-1} w = g \quad \text{if } \Delta_v g \equiv \nabla \cdot (v \nabla g) = -w,$$

and $g = 0$ at the free boundary of v. Then integrating by parts yields the identity

$$\int v|\nabla(-\Delta_v)^{-1}v_\tau|^2 = \tfrac{\mathrm{d}}{\mathrm{d}\tau}\left[-\tfrac{1}{2}\int|\nabla v|^2 - \tfrac{1}{2(4+N)}\int v|y|^2\right] + J, \tag{3.5}$$

where J corresponds to the exponentially small term,

$$J = \mathrm{e}^{-\gamma\tau}\int v^p(-\Delta_v)^{-1}v_\tau. \tag{3.6}$$

Integrating (3.5) over $(0, T)$ yields

$$\int_0^T\int v|\nabla(-\Delta_v)^{-1}v_\tau|^2 + \tfrac{1}{2}\int|\nabla v(T)|^2 + \tfrac{1}{2(4+N)}\int v(T)|y|^2 \le C + \int_0^T J,$$

so that, if the exponential term (3.6) $J \in L^1(\mathbb{R}_+)$, this yields extra uniform estimates,

$$\sqrt{v}\,\nabla(-\Delta_v)^{-1}v_\tau \in L^2(\mathbb{R}\times\mathbb{R}_+) \;\text{ and }\; \nabla v,\ \sqrt{v}|y| \in L^\infty(\mathbb{R}_+; L^2).$$

Note that, obviously, (3.5) does not imply existence of a Lyapunov function (the nonautonomous PDE (3.4) is not a gradient system). Anyway, since (3.5) gives a rather strong estimate of v_τ for $\tau \gg 1$, this makes it possible to pass to the limit $\tau \to \infty$ and establish stabilization to an equilibrium point (see the technique in [26, pp. 116–117]), which is unique by the obvious mass-monotonicity with time of the solution.

The symmetry of the Fréchet derivative (2.12) at F looks like a certain "remnant" of the fact that the original PDE is a gradient system.

4. Center subspace patterns for the $2m$th-order TFE.

We consider the $2m$th-order TFE with absorption (1.24) with the critical absorption (Fujita) exponent (1.25). The proper setting of a standard "zero contact angle" FBP for the TFE includes $m + 1$ free boundary conditions at the free boundary $\Gamma_0 = \partial\Omega(t)\times\mathbb{R}_+$ ($\Omega(t)$ is the support of $u(\cdot, t)$ at time $t > 0$),

$$u = \nabla u = \cdots = \tfrac{\partial^{m-1}u}{\partial\nu^{m-1}} = \nu\cdot\nabla(u^n\Delta^{m-1}u) = 0, \tag{4.1}$$

where ν is the unit outward normal to $\partial\Omega(t)$ that is assumed to be sufficiently smooth.

4.1. Similarity solutions.

The similarity solutions of the pure TFE

$$u_t = (-1)^{m+1}\nabla\cdot(u^n\nabla\Delta^{m-1}u) \tag{4.2}$$

take the standard form (1.19) with

$$\beta = \tfrac{1}{2m+nN}. \tag{4.3}$$

One can see that the critical exponent (1.25) is precisely the one for which the PDE (1.24) possesses the same group of scaling transformations. Then the rescaled profile F satisfies the radial restriction of the $2m$th-order elliptic equation

$$\mathbf{A}(F) = (-1)^{m+1}\nabla\cdot(F^n\nabla\Delta^{m-1}F) + \beta\nabla F\cdot y + \beta NF = 0. \tag{4.4}$$

It seems that, for any $m \ge 3$, the questions of existence and uniqueness of a solution $F(y) > 0$ in B_1 remain open. It is clear that, for large m, a standard approach to existence based on a multiparametric shooting leads to a complicated geometric analysis (though some general conclusions in this geometry are likely). We expect that the approach based on the n-branching (or a continuous homotopy connection with

$n=0$) via the classical theory [43] makes it possible to explain properties solutions, at least, for small $n>0$ by branching from the linear case $n=0$ (but, surely, a standard approach to smooth branching does not apply). For the Cauchy problem, the spectral and other properties of the corresponding linear operator (4.4) for $n=0$ are given in [13] and can be used to clarify the behavior for small $n>0$. For the FBP (4.1), an extra analysis of the linearized elliptic PDE is necessary.

As usual, the case $n=1$ provides us with the explicit solution. Writing the ODE (4.4) in the radial divergent form (here y is actually $|y|$)

$$(y^{N-1}F(\Delta^{m-1}F)')' = (-1)^m\beta(y^N F)',$$

on integration we obtain $\Delta^{m-1}F = (-1)^m \frac{1}{2}\beta y^2$. Integrating this linear ODE $2m-2$ times yields the positive solution in B_1

$$F(y) = c_0(1-|y|^2)^m, \quad \text{where } c_0 = \tfrac{1}{2}\,\tfrac{N!!}{(2m)!!(2m+N)!!}. \tag{4.5}$$

4.2. Linearized operator. We next follow the same scheme of the asymptotic analysis as in section 2. Similar to (2.10), we introduce the linearized operator

$$\begin{aligned}\mathbf{A}'(F)Y &= (-1)^{m+1}\nabla\cdot(F^n\nabla\Delta^{m-1}Y)\\ &\quad + (-1)^{m+1}\nabla\cdot(nF^{n-1}Y\nabla\Delta^{m-1}F) + \beta\nabla Y\cdot y + \beta NY.\end{aligned} \tag{4.6}$$

Using the ODE (4.4), we have that

$$(-1)^{m+1}\nabla\cdot(nF^{n-1}\nabla\Delta^{m-1}F) = -\beta nN, \;\; (-1)^{m+1}nF^{n-1}\nabla\Delta^{m-1}F = -\beta ny,$$

so (4.6) can be written in the form

$$\mathbf{A}'(F)Y = (-1)^{m+1}\nabla\cdot(F^n\nabla\Delta^{m-1}Y) + \beta N(1-n)y\cdot\nabla Y + \beta N(1-n)Y, \tag{4.7}$$

and we again observe that $n=1$ is a special case.

4.3. The self-adjoint case $n=1$. Plugging the profile (4.5) into (4.7) yields the following symmetric form of the operator:

$$\begin{aligned}\mathbf{A}'(F)Y &= c_0(-1)^{m+1}\nabla\cdot((1-|y|^2)^m\nabla\Delta^{m-1}Y)\\ &\equiv c_0(-1)^{m+1}[D^m((1-|y|^2)D^mY) + m(m-1)N\Delta^{m-1}Y],\end{aligned} \tag{4.8}$$

where D^m denotes $\Delta^{m/2}$ for m even and $\nabla\Delta^{(m-1)/2}$ for m odd. For instance, for $N=1$ and $m=3$, we have

$$\mathbf{A}'(F)Y = c_0(1-y^2)^2[((1-y^2)Y''')''' + 6Y^{(4)}].$$

Having the symmetric operator (4.8) in C_0^∞, we next determine its self-adjoint extensions [10]. In particular, there exists the extension with discrete spectrum and polynomial eigenfunctions in the radial setting (the nonradial case is covered by using the spherical polynomials as in (2.17)). The eigenvalues λ_k for the polynomials $\psi_k(y)$ given in (2.19) are calculated by using (4.8),

$$\lambda_k = -c_0(k+2)k\ldots[k+2-2(m-2)](k+N+2)(k+N)\ldots[k+N-2(m-2)] \tag{4.9}$$

for $k=2(m-3), 2(m-2), \ldots$. Using the eigenfunction expansion in terms of a complete and closed subset of polynomials $\{\psi_k\}$ partially justifies the asymptotic center subspace analysis of the corresponding rescaled equation (2.9), which yields the same ODE (2.24) and hence the asymptotics (2.26). Here in the critical case (1.25) we still have $\frac{1}{p-1} = \beta N$ with β given by (4.3). Finally, we arrive at the asymptotic pattern (1.22), where 4 is replaced by $2m$.

4.4. The general case $n \neq 1$. We do not have such a self-adjoint operator, but anyway, once $F > 0$ in B_1 is determined, we obtain the radial eigenfunction ψ_0 for $\lambda_0 = 0$ from the scaling symmetry group (2.28) (the exponent $\frac{4}{n}$ is replaced by $\frac{2m}{n}$) of (4.4). We can also guarantee that (4.7) has compact resolvent provided that $n > 0$ is not large, so $\lambda_0 = 0$ is an isolated eigenvalue. The rest of the center subspace behavior via the expansion (2.33) remains unchanged and leads to similar logarithmically perturbed asymptotic patterns. A rigorous justification is a hard open problem.

5. Logarithmically perturbed patterns in the Cauchy problem.

The asymptotic behavior and similarity solutions for the TFE (1.12) or (1.24) posed in the whole space $\mathbb{R}^N \times \mathbb{R}_+$ are less studied in the literature. For $n \in (0, \frac{3}{2})$, in the Cauchy problem, the solutions exhibiting the "maximal regularity" at the interfaces are oscillatory and of changing sign. See [17, 18] and the book [25, Ch. 1] for the correct meaning of the Cauchy problem for TFEs and further examples. For such solutions, we need to assume that u^n in (4.1) is replaced by $|u|^n$. Therefore, from now on, in all the expressions and equations we use the convention that

$$\begin{gathered} u^n, f^n, v^n, w^n, \dots \text{ are replaced by } |u|^n, |f|^n, |v|^n, |w|^n, \dots \text{ and} \\ u^p, f^p, v^p, w^p, \dots \text{ are replaced by } |u|^{p-1}u, |f|^{p-1}f, |v|^{p-1}v, |w|^{p-1}w, \dots . \end{gathered} \tag{5.1}$$

We must admit that solutions of changing sign are less relevant for many known physical applications of TFEs. Nevertheless, for general PDE theory, it is key and of principal importance to include the Cauchy problem and to show that the basic techniques developed above apply to these much more complicated oscillatory solutions.

The idea of sign changing solutions of TFEs is straightforward. Indeed, the oscillatory properties of such solutions are a manifestation of the fact that TFEs (4.2) are "homotopic," i.e., can be continuously deformed (e.g., as $n \to 0$) via nonsingular uniformly parabolic PDEs with analytic coefficients (for details see [18, section 14]) to the linear *polyharmonic equation*

$$u_t = (-1)^{m+1}\Delta^m u \quad \text{in } \mathbb{R}^N \times \mathbb{R}_+. \tag{5.2}$$

By classical parabolic theory (see, e.g., Èĭdel'man [14]), given initial data $u_0 \in L^1$, there exists the unique solution of the Cauchy problem for (5.2) defined by the convolution

$$u(x,t) = b(x,t) * u_0, \quad b(x,t) = t^{-\frac{N}{2m}} F(y), \;\; y = x/t^{\frac{1}{2m}}, \tag{5.3}$$

where $b(x,t)$ is the fundamental solution of the operator $D_t - (-1)^{m+1}\Delta^m$. For any $m \geq 2$, the rescaled kernel $F = F(|y|)$ is oscillatory as $y \to \infty$, so this property of changing sign is inherited by L^1 solutions of (5.2). Assuming a continuous (homotopic) deformation of a class of solutions of (1.12) as $n \to 0^+$, this confirms that the TFE admits oscillatory solutions of changing sign at least for not very large $n > 0$. Continuity and homotopy concepts are effective for treating the Cauchy problem for higher-order TFEs; see other examples in [18].

Then the source-type solutions of the TFE take the same form (1.19), where the radial function F of changing sign solves the ODE (1.20) with the convention (5.1). We begin with the linear case $n = 0$, which by continuity is going to describe some properties of source-type solutions for sufficiently small $n > 0$.

5.1. Properties of the rescaled fundamental solution for $n = 0$. The linear ODE

$$\mathbf{A}(F) \equiv -\Delta^2 F + \tfrac{1}{4}\nabla F \cdot y + \tfrac{N}{4} F = 0 \quad \text{in } \mathbb{R}^N \tag{5.4}$$

is precisely the elliptic equation for the rescaled kernel F of the fundamental solution in (5.3). Therefore, the similarity profile $F(y)$ exists and is unique under the assumption

$$\textstyle\int F(y)\,\mathrm{d}y = 1 \tag{5.5}$$

(in view of existence-uniqueness of the fundamental solution).

Let us next describe an important relation between similarity profiles for the FBP and the Cauchy problem. Without loss of generality, we consider the case $N = 1$, where on integration once (5.4) takes the form

$$F''' = \tfrac{1}{4} F y. \tag{5.6}$$

It is easy to find all decaying profiles corresponding to the Cauchy problem with the exponential WKBJ asymptotics as $y \to +\infty$,

$$F(y) \sim y^{-\frac{1}{3}} \mathrm{e}^{a y^{4/3}}, \quad \text{with } a \text{ satisfying } a^3 = \tfrac{1}{4}\left(\tfrac{3}{4}\right)^3. \tag{5.7}$$

There exist two complex conjugate roots for exponentially decaying profiles

$$a_\pm = -\tfrac{3}{8}\, 4^{-\frac{1}{3}}(1 \pm \mathrm{i}\sqrt{3}) \equiv -c_1 \pm \mathrm{i} c_2. \tag{5.8}$$

This yields a *two-dimensional* bundle of oscillatory solutions with the behavior

$$F(y) \sim y^{-\frac{1}{3}} \mathrm{e}^{-c_1 y^{4/3}} \left[A_1 \cos\left(c_2 y^{\frac{4}{3}}\right) + A_2 \sin\left(c_2 y^{\frac{4}{3}}\right)\right] \quad \text{as } y \to \infty, \tag{5.9}$$

where A_1 and A_2 are arbitrary constants. The algebraic factor $y^{-1/3}$ is obtained by a standard asymptotic WKBJ method. We observe here the periodic behavior with a *single* fundamental frequency (a result we will refer to in the TFE analysis below).

PROPOSITION 5.1. *For $N = 1$, the rescaled profile of the Cauchy problem $F = F_\infty$ given by* (5.4), (5.5) *is the limit of FBP similarity profiles on bounded intervals,*

$$F_\infty = \lim F_k, \tag{5.10}$$

where each $F_k(y)$ is defined on interval $(-y_k, y_k)$,

$$F_k(\pm y_k) = F_k'(\pm y_k) = 0, \quad \textit{and} \tag{5.11}$$

$$y_k = \left(\tfrac{\pi}{c_2}\, k\right)^{\frac{3}{4}} (1 + o(1)) \quad \textit{as } k \to \infty. \tag{5.12}$$

Proof. The geometric aspect of such a property is obvious in view of the oscillatory behavior in (5.9). The convergence as $k \to \infty$ follows from straightforward computations related to the whole exponential bundle including (5.9) and the growing counterpart

$$F(y) = y^{-\frac{1}{3}} \mathrm{e}^{a_0 y^{4/3}} + \cdots, \quad \text{with } a_0 = \frac{3}{4}\, 4^{-\frac{1}{3}}.$$

Then solving the FBP problem (5.11) yields the asymptotic equality $\cos(c_2 y_k^{4/3} + \text{const}) = 0$, whence the asymptotics (5.12). □

We also expect the following *Sturm property* be valid:

$$F_k(y) \quad \text{has precisely } k \text{ zeros on } (0, y_k). \tag{5.13}$$

Such a zero-number property is easily seen for $k \gg 1$ but is not obvious for smaller k's.

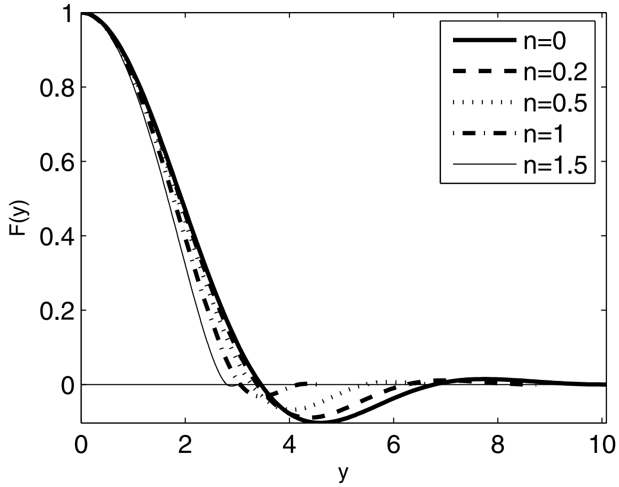


FIG. 2. *The oscillatory Cauchy problem profiles satisfying* (5.14). *Parameters of shooting are* $F''(0) = -0.3379890$ *(*$n = 0$*)*, -0.3414702 *(*$n = 0.2$*)*, -0.3490986 *(*$n = 0.5$*)*, -0.3697143 *(*$n = 1$*)*, *and* -0.4052680 *(*$n = 1.5$*)*.

5.2. Similarity profiles for $n > 0$: Existence and uniqueness.

PROPOSITION 5.2. *For* $N = 1$ *and* $n \in (0,1)$*, the ODE* (1.20), (5.1) *in* $\mathbb{R}$ *admits a unique solution* $F \in C^3$ *of unit mass. The solution* $F(y)$ *is symmetric, compactly supported, and oscillatory near finite interfaces at* $y = \pm y_0$.

Proof. For $N = 1$ the ODE (1.20) has the form

$$|F|^n F''' = \beta F y, \quad y \in \mathbb{R}. \tag{5.14}$$

Dividing by $|F|^n$ and setting $|F|^{-n}F = g$ yields

$$(|g|^\alpha g)''' = \beta g y, \quad y \in \mathbb{R}, \quad \alpha = \tfrac{n}{1-n}. \tag{5.15}$$

Then existence and uniqueness of a compactly supported solution $F \in C^3$ for any $n \in (0,1)$ follow from the results in Bernis and McLeod [6]. □

For $n \in [1, \frac{3}{2})$ solutions of (5.14) are less regular (see below), so the techniques in [6] do not apply directly, but we expect that the existence-uniqueness result remains valid and can be extended further to some interval $n \in [\frac{3}{2}, n_{\rm h})$; see below.

In Figure 2 we have shown these similarity profiles for some $n > 0$ including the linear case $n = 0$ leading to the ODE (5.6) for the fundamental rescaled profile. Here we observe convergence of the fundamental profiles as $n \to 0^+$, which is justified rigorously if all the zeros are "transversal" and isolated except the last one; see below.

5.3. Oscillatory properties via periodic orbits. We next describe the oscillatory properties of such sign changing profiles $F(y)$ near interfaces. We rescale F to have that

$$\operatorname{supp} F = [-1, 1].$$

It was shown in [17] that the asymptotic behavior of $F(y)$ satisfying (5.14) near the interface point $y \to 1^-$ is given by the expansion

$$F(y) = (1-y)^\mu \phi(s), \quad s = \ln(1-y), \quad \mu = \tfrac{3}{n}, \tag{5.16}$$

where, after scaling $\phi \mapsto \beta^{\frac{1}{n}}\phi$, the oscillatory component ϕ satisfies the following autonomous ODE (we omit exponentially small terms):

$$\phi''' + 3(\mu-1)\phi'' + (3\mu^2 - 6\mu + 2)\phi' + \mu(\mu-1)(\mu-2)\phi + \tfrac{\phi}{|\phi|^n} = 0. \tag{5.17}$$

Oscillatory periodic orbits: Existence. We are now interested in periodic solutions $\phi_*(s)$ of (5.17), which, according to (5.16), can determine the simplest typical (and possibly stable and generic) oscillatory behavior of solutions near interfaces when $s = \ln(1-y) \to -\infty$ as $y \to 1^-$. There are several classic methods of ODE theory for establishing existence and multiplicity of periodic solutions of finite-dimensional dynamical systems. These are various topological techniques, such as rotations of vector fields, index, and degree theory; see [33, sections 13, 14]. Another approach is based on branching theory [43, Ch. 6]. In our case, such an n-branching approach is especially effective since for $n = 0$ the unique solution F is the rescaled kernel of the fundamental solution (a rigorous justification of some aspects of branching for such degenerate equations can be a hard problem). We also mention papers [44, 35, 32] containing further related references and methods concerning modern theory of periodic solutions of higher-order nonlinear ODEs. In general, equations like (5.17) are difficult to study; particularly, the main difficulty is proving *uniqueness* of such periodic orbits. Therefore, later on, together with analytic techniques, we will need also to rely on careful numerical evidence on existence, uniqueness, and stability of periodic solutions.

It is curious that for $n = 1$, the unique periodic solution can be detected by a direct algebraic approach; see [17, section 7.4].

Proposition 5.3. *For $n = 1$, the ODE* (5.17) *has a unique T-periodic solution, with*

$$T = -2\ln s > \theta = 1.9248\ldots, \tag{5.18}$$

where $\theta = 0.381966\ldots$ is the unique root on the interval $(0, 1)$ *of the cubic equation*

$$\theta^3 - 2\theta^2 - 2\theta + 1 = 0. \tag{5.19}$$

Indeed, for $n = 1$, the nonlinearity in (5.17) is $\operatorname{sign}\phi$ and the ODE is linear in the positivity and negativity domain of solutions,

$$\phi''' + 6\phi'' + 11\phi' + 6\phi \pm 1 = 0,$$

and so can be solved explicitly. Matching positive and negative branches leads to the result.

Let us now state the main result concerning periodic orbits of the ODE (5.17).

Theorem 5.4. *The ODE* (5.17) *admits a nontrivial stable periodic solution $\phi_*(s)$ of changing sign for all*

$$0 < n < n_{\rm h} \in (\tfrac{3}{2}, n_+), \quad \textit{where} \quad n_+ = \tfrac{9}{3+\sqrt{3}} = 1.9019238\ldots. \tag{5.20}$$

Uniqueness of such periodic $\phi_*(s)$ in the interval (5.20) is still open.

Proof. For the interval

$$0 < n < \tfrac{3}{2}, \tag{5.21}$$

the proof of existence is performed in [17, p. 292] by a shooting argument. Numerical representation of periodic solutions is given therein on page 294; see also [25, p. 143]. We need to point out the main two ingredients of the proof in [17]:

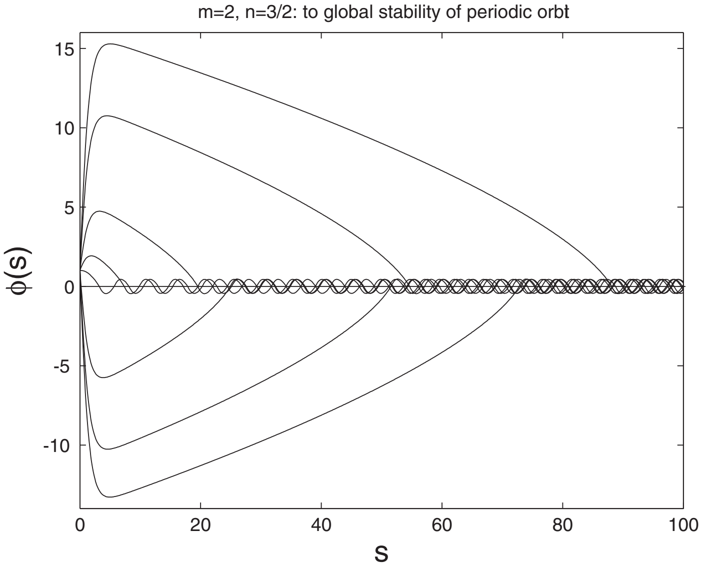


FIG. 3. *Convergence to the stable periodic solutions of* (5.17) *for* $n = \frac{3}{2}$ *for various Cauchy data posed at* $s = 0$.

(i) it is shown that for exponents (5.21) no orbits of the dynamical system (5.17) are attracted to infinity as $s \to +\infty$; i.e., all orbits stay uniformly bounded; and

(ii) as a consequence, then (5.17) is a dissipative dynamical system having a bounded absorbing set.

Dissipative dynamical systems are known to admit periodic solutions in a rather general setting [33, section 39] provided these are nonautonomous (so the period is fixed). For the autonomous system (5.17), the proof in [17, section 7.1] was completed by shooting. Note that, in view of the last term, (5.17) is not a smooth dynamical system and solutions are not locally C^3-smooth. Nevertheless, as local analysis shows [17, p. 291], at least for $n \in (0, 2)$, the nonlinearity is integrable to guarantee local extensions of solutions through generic “transversal” zeros. This means that the equivalent integral equation is well-posed and is composed from compact operators in a certain topology (this is necessary for application of classic methods of branching in Banach spaces [43, Ch. 7]). We continue to deal with the differential equation, where the justification of calculus is done by local analysis.

It turns out that both properties (i) and (ii) also remain valid for $n = \frac{3}{2}$, so that a periodic solution ϕ_* also exists and is stable; see Figure 3. For the extension of ϕ_* to $n > \frac{3}{2}$, we will use the following crucial stability result.

PROPOSITION 5.5. *If the periodic solution* $\phi_*(s)$ *of* (5.17) *persists for all* $\frac{3}{2} \le n_{\rm h} < 3$, *then it is stable and hyperbolic on this interval.*

Proof. Note that, for $n \in (\frac{3}{2}, 3)$, there exist two unstable constant equilibria of equation (5.17),

$$\phi_\pm = \pm \big[-\tfrac{1}{\mu(\mu-1)(\mu-2)} \big]^{\frac{1}{n}} \quad \text{for } \ n \in (\tfrac{3}{2}, 3), \tag{5.22}$$

and we expect a stable periodic motion in between. Consider the eigenvalue problem for the ODE (5.17) linearized about the T-periodic solution ϕ_* by setting $\phi = \phi_* + Y$,

$$Y''' + 3(\mu-1)Y'' + (3\mu^2 - 6\mu + 2)Y' + \mu(\mu-1)(\mu-2)Y + (1-n)|\phi_*|^{-n}Y = \lambda Y.$$

As usual, assuming that $\lambda \in \mathbb{C}$, multiplying this by the complex conjugate $\overline{Y}$ in $L^2(0,T)$, taking the conjugate and multiplying by Y, and summing up both yields (5.23)

$$-3(\mu-1)\int |Y'|^2 + \mu(\mu-1)(\mu-2)\int |Y|^2 + (1-n)\int |\phi_*|^{-n}|Y|^2 = \tfrac{\lambda+\bar{\lambda}}{2}\int |Y|^2.$$

Since all three terms on the left-hand side of (5.23) are negative for any $\frac{3}{2} < n < 3$, the result follows. The case $n = \frac{3}{2}$ is similar since just the second term vanishes. □

Thus, by classic branching theory [43, Ch. 6], stable hyperbolic periodic solutions are locally extensible relative to the parameter $n \ge \frac{3}{2}$. In particular, using the hyperbolicity of ϕ_* for $n = \frac{3}{2}$, we conclude that the periodic solution exists in an interval $n \in [\frac{3}{2}, \frac{3}{2} + \delta)$ with some $\delta > 0$, and the interval of existence must be open from the right-hand side.

Finally, let us justify the estimate in (5.20). To this end, we multiply (5.17) by ϕ_*' and integrate over $(0,T)$ to get for any $n \in (0,2)$

$$-\int (\phi_*'')^2 + (3\mu^2 - 6\mu + 2)\int (\phi_*')^2 = 0,$$

so that one needs

$$3\mu^2 - 6\mu + 2 > 0 \quad \Longrightarrow \quad \mu = \tfrac{3}{n} > \mu_+ = \tfrac{3}{n_+} = \tfrac{3+\sqrt{3}}{3}.$$

This completes the proof of Theorem 5.4. □

On heteroclinic bifurcation. Since the periodic orbit $\phi_*(s)$ remains stable and hyperbolic in the whole interval of existence (5.20), the end point $n = n_{\rm h}$ cannot be any kind of subcritical saddle-node bifurcation, at which two branches meet each other. Classic bifurcation and branching theory [33, 43] then suggests that at $n = n_{\rm h}^-$ the dynamical system (5.17) undergoes a *heteroclinic bifurcation* when the period increases without bound (this claim needs further study and a full analytical justification); see standard scenarios in Perko [41, Ch. 4]. Note that, by Proposition 5.5, the heteroclinic orbit that occurred remains stable and hyperbolic.

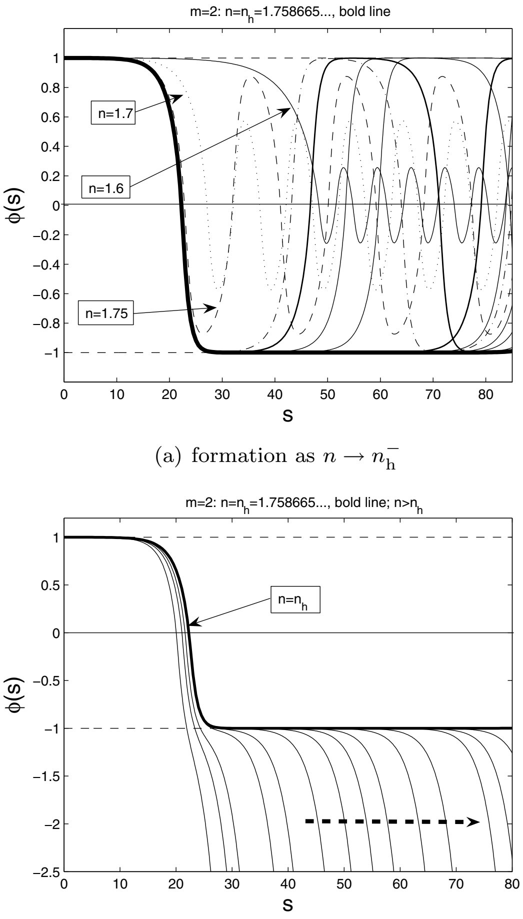
Numerically, $n_{\rm h}$ is given by

$$n_{\rm h} = 1.7598665026\ldots. \tag{5.24}$$

Figure 4 shows the formation of the heteroclinic orbit in both limits: as $n \to n_{\rm h}^-$ (a) and $n \to n_{\rm h}^+$ (b). This bifurcation exponent $n_{\rm h}$ plays an important role and shows the parameter range of n's, for which many ODE profiles near interfaces are *oscillatory*, except those that approach the interface point $s = -\infty$, the stable manifold of the constant equilibrium (5.22). In the interval (5.21), this manifold of orbits of constant sign is empty, so that all the orbits near $s = -\infty$ are oscillatory and coincide with the periodic one $\phi_*(s + s_0)$, where $s_0 \in \mathbb{R}$ is a parameter of shifting. Indeed, this also characterizes important oscillatory features of the PDE. Note that some kind of a "heteroclinic bifurcation" phenomenon also exists for the sixth-order ($m = 3$) and higher-order TFEs with more difficult mathematics involved; see [18, section 13] and [25, pp. 142–147].

On one-dimensional shooting for $n \in (1, n_{\rm h})$. As a key application of the above oscillation analysis, we have that, according to (5.16), for all $n \in (0, n_{\rm h})$, there exists a one-dimensional bundle of oscillatory orbits of changing sign

$$F(y) = (1-y)^{\frac{3}{n}}\phi_*(\ln(1-y) + s_0) + \ldots, \tag{5.25}$$



(a) formation as $n \to n_{\mathrm{h}}^-$

(b) formation as $n \to n_{\mathrm{h}}^+$

FIG. 4. *Formation of a heteroclinic orbit as $n \to n_{\mathrm{h}}$.*

where $s_0 \in \mathbb{R}$ is an arbitrary parameter of phase shift in the periodic orbit $\phi_*(s)$. Recall that, for the ODE (5.14), we need to shoot just a single symmetry condition at the origin,

$$F'(0) = 0 \quad (F(0) \neq 0), \tag{5.26}$$

so the one-dimensional bundle (5.25) is well suited for this. In view of oscillatory character of the behavior in (5.25), it is not difficult to prove the existence of such an s_0 to satisfy (5.26), while uniqueness (as expected) remains open.

Further comments about n_{h}. For any $n > n_{\mathrm{h}}$, the behavior in the ODE (5.17) becomes exponentially unstable, and we did not observe oscillatory or changing sign patterns. This suggests that precisely above $n = n_{\mathrm{h}}$, the ODE (and the corresponding PDE) loses its natural similarities with the linear one for $n = 0$ (though a continuous homotopic connection is expected to be still available; i.e., some local properties of solutions dramatically change at n_{h}).

Thus, in the range $n \in (\frac{3}{2}, 3)$, (5.17) possesses the positive constant solution $\phi(s) \equiv \phi_+$ given in (5.22). This gives the behavior (2.35), so that, for such solutions, formally, the FBP and the Cauchy problem may coincide in the ODE setting. But this is not the case for all the solutions since for $n \in (\frac{3}{2}, n_{\rm h})$ there are other oscillatory profiles with a similar (actually, a bit less) regularity at the interfaces, so that the Cauchy problem demands oscillatory solutions, while the FBP can admit positive solutions; for more details see [17, section 9]. In the parameter range $n \geq n_{\rm h}$, the oscillatory behavior is no longer generic, so we expect a certain improvement of the positivity preserving properties of the TFE, where the Cauchy problem and the FBP may coincide; see further discussion in [17, section 9.4].

5.4. The TFE with critical absorption $\boldsymbol{p = p_0}$. The formal asymptotics for the TFE (1.5), (1.6) is now calculated similarly using the center subspace spanned by the eigenfunction (2.29). Of course, we then do not gain any explicit mathematics or symmetric operators as for $n = 1$ in the case of the FBP.

The main ideas of the analysis can be extended to the $2m$th-order case, where many aspects of source-type and general solutions of the Cauchy problem for the TFEs remain mathematically open. The oscillatory character of solutions near the interface for $m = 3$ was studied in [18, section 13]; see also [25, section 3.7] for further examples for $m \geq 3$ and other oscillatory PDEs.

5.5. Supercritical range $\boldsymbol{p > p_0}$. We use the same scaling (3.1) and obtain the exponentially perturbed rescaled PDE (3.2), which suggests that the solutions behave as $t \to \infty$ as the source-type solution with a finite positive mass attained at $\tau = +\infty$ (no proof is yet available).

Appendix. The linearized operator is not symmetric when $\boldsymbol{n \neq 1}$. We prove that, in the FBP setting, the linearized operator (2.11) admits a self-adjoint extension only when $n = 1$. Without loss of generality we consider the one-dimensional case, and we formulate first the following results we are already familiar with.

Proposition A.1. *The linearized operator* (2.11) *in* $\mathbb{R}$ *is symmetric in some weighted space* L^2_ρ *when* $n = 1$.

Proof. For $N = n = 1$, the linearized operator is given by

$$\mathbf{A}'(f)Y = -(fY''')' - (Yf''')' + \tfrac{1}{5}\,(Yy)'. \tag{A.1}$$

For this to be symmetric in L^2_ρ with some weight $\rho \geq 0$, we require that [37, section 1]

$$\mathbf{A}'(f)Y \equiv \tfrac{1}{\rho}\,\left[(p_0Y'')'' - (p_1Y')' + p_2Y\right]. \tag{A.2}$$

Expanding the right-hand sides of these equations and comparing coefficients yields the following system:

$$Y'''' : \ -f = \tfrac{p_0}{\rho}, \tag{A.3}$$

$$Y''' : \ -f' = \tfrac{2p_0'}{\rho}, \tag{A.4}$$

$$Y'' : \ 0 = \tfrac{p_0''-p_1}{\rho}, \tag{A.5}$$

$$Y' : \ -f''' + \tfrac{1}{5}y = -\tfrac{p_1'}{\rho}, \tag{A.6}$$

$$Y : \ -f'''' + \tfrac{1}{5} = \tfrac{p_2}{\rho}. \tag{A.7}$$

We know the exact solution of the ODE for f when $n = 1$ (see (1.21)):

$$f(y) = \tfrac{1}{120}\,(a^2 - y^2)^2 \ \text{ for } y \in (-a, a). \tag{A.8}$$

Substituting this into (A.7) yields $p_2 = 0$. Equation (A.6) yields $p_1 = C$, where C is a constant. Equations (A.3) and (A.4) yield $p_0^2 = f$ and $\rho = -f^{-1/2}$. Equation (A.5) is thus the consistency condition and is satisfied since it yields $p_1 = C$ (since $p_0'' = p_1 = C$). Thus the linearized operator for the TFE is symmetric if $n = 1$. □

THEOREM A.2. *For $N = 1$ and $n \neq 1$, operator* (2.11) *is not symmetric in L^2_ρ for any weight $\rho > 0$.*

Proof. The ODE for $f > 0$ for any $n > 0$ is

$$-(f^n f''')' + \tfrac{1}{n+4}\,(fy)' = 0. \tag{A.9}$$

The linearized operator (2.11) is given by

$$\mathbf{A}'(f)Y = -(f^n Y''')' - n(f^{n-1} Y f''')' + \tfrac{1}{n+4}\,(Yy)'. \tag{A.10}$$

For this to be symmetric, we require identity (A.2) to hold. Comparing coefficients yields

$$Y'''' : \quad -f^n = \tfrac{p_0}{\rho}, \tag{A.11}$$

$$Y''' \;:\quad -nf^{n-1}f' = \tfrac{2p_0'}{\rho}, \tag{A.12}$$

$$Y'' \;:\quad 0 = \tfrac{p_0''-p_1}{\rho}, \tag{A.13}$$

$$Y' \;:\quad -nf^{n-1}f''' + \tfrac{y}{n+4} = -\tfrac{p_1'}{\rho}, \tag{A.14}$$

$$Y \;:\quad -n(n-1)f^{n-2}f''' - nf^{n-1}f'''' + \tfrac{1}{n+4}. \tag{A.15}$$

From this

$$p_0^2 = f^n,\;\; p_1 = p_0'',\;\; \rho = -f^{-n/2},\;\; p_2 = \rho\big[-n(n-1)f^{n-2}f''' - nf^{n-1}f'''' + \tfrac{1}{n+4}\big],$$

and the consistency condition is

$$f^{\frac{n}{2}}(f^{\frac{n}{2}})''' = -nf^{n-1}f''' + \tfrac{1}{n+4}\,y. \tag{A.16}$$

To see if this coincides with (A.9) for some f, we use a Taylor expansion of $f(y)$ and check if (A.16) and (A.9) produce the same coefficients for f. To do this we set $f(0) = 1,\; f'(0) = f'''(0) = 0$, and $f''(0) = b \in \mathbb{R} \setminus \{0\}$, differentiate (A.16) and (A.9) the required number of times, and set $y = 0$. The expansions coincide up to the coefficient of y^3, but the coefficients of y^4 coincide only if

$$b = \pm \tfrac{\sqrt{-6n(n^2+2n-8)(3n-2)}}{3n^3+6n^2-24n}. \tag{A.17}$$

Since we require $b \in \mathbb{R} \setminus \{0\}$, we must have $n \in (-4, 0) \cup (\frac{2}{3}, 2)$. This gives us a range of values of n, for which the linearized operator may be symmetric. To check whether it is, we examine the coefficient of y^6 for (A.16) and (A.9). If the operator is symmetric, then the same value of b should be obtained as in the coefficients of y^4 for both equations. For the coefficients of y^6 to coincide, we require

$$b = 0, \;\text{ or }\; b = \pm \tfrac{2\sqrt{2}\sqrt{n(9n^3-40n^2-188n+464)(3n-2)}}{9n^4-40n^3-188n^2+464n}\,, \tag{A.18}$$

and since we require $b \in \mathbb{R} \setminus \{0\}$, we discard $b = 0$. For this b to coincide with (A.17), we require $n = \frac{2}{3}$. This contradicts the fact that we must have $n \in (-4, 0) \cup (\frac{2}{3}, 2)$ for the linearized operator (2.11) to have a chance of being symmetric and admit a suitable (Friedrichs) self-adjoint extension. Hence the linearized operator is not symmetric if $n \neq 1$. □

Acknowledgments. The authors would like to thank J. D. Evans for discussions on thin film models with nonconservative aspects and A. Leger for efficient consulting with the authors on numerical methods for higher-order ODEs.

REFERENCES

[1] J. Becker and G. Grün, *The thin-film equation: Recent advances and some new perspectives*, J. Phys. Condens. Matter, 17 (2005), pp. S291–S307.

[2] D. J. Benney, *Long waves on liquid films*, J. Math. Phys., 45 (1966), pp. 150–155.

[3] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[4] F. Bernis, J. Hulshof, and J. R. King, *Dipoles and similarity solutions of the thin film equation in the half-line*, Nonlinearity, 13 (2000), pp. 413–439.

[5] F. Bernis, J. Hulshof, and F. Quirós, *The "linear" limit of thin film flows as an obstacle-type free boundary problem*, SIAM J. Appl. Math., 61 (2000), pp. 1062–1079.

[6] F. Bernis and J. B. McLeod, *Similarity solutions of a higher order nonlinear diffusion equation*, Nonlinear Anal., 17 (1991), pp. 1039–1068.

[7] F. Bernis, L. A. Peletier, and S. M. Williams, *Source type solutions of a fourth order nonlinear degenerate parabolic equation*, Nonlinear Anal., 18 (1992), pp. 217–234.

[8] A. J. Bernoff and T. P. Witelski, *Linear stability of source-type similarity solutions of the thin film equation*, Appl. Math. Lett., 15 (2002), pp. 599–606.

[9] A. L. Bertozzi and M. C. Pugh, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 41 (1998), pp. 625–651.

[10] M. S. Birman and M. Z. Solomjak, *Spectral Theory of Self-Adjoint Operators in Hilbert Space*, D. Reidel, Dordrecht, Tokyo, 1987.

[11] M. Bowen, J. Hulshof, and J. R. King, *Anomalous exponents and dipole solutions for the thin film equation*, SIAM J. Appl. Math., 62 (2001), pp. 149–179.

[12] J. A. Carrillo and G. Toscani, *Long-time asymptotic behaviour for strong solutions of the thin film equations*, Comm. Math. Phys., 225 (2002), pp. 551–571.

[13] Yu. V. Egorov, V. A. Galaktionov, V. A. Kondratiev, and S. I. Pohozaev, *Asymptotic behaviour of global solutions to higher-order semilinear parabolic equations in the supercritical range*, Adv. Differential Equations, 9 (2004), pp. 1009–1038.

[14] S. D. Èĭdel'man, *Parabolic Systems*, North–Holland, Amsterdam, London, 1969.

[15] C. M. Elliott and H. Garcke, *On the Cahn–Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[16] C. Elliott and Z. Songmu, *On the Cahn-Hilliard equation*, Arch. Ration. Mech. Anal., 96 (1986), pp. 339–357.

[17] J. D. Evans, V. A. Galaktionov, and J. R. King, *Source-type solutions of the fourth-order unstable thin film equation*, European J. Appl. Math., 18 (2007), pp. 273–321.

[18] J. D. Evans, V. A. Galaktionov, and J. R. King, *Unstable sixth-order thin film equation.* I. *Blow-up similarity solutions*, Nonlinearity, 20 (2007), pp. 1799–1841.

[19] J. D. Evans, V. A. Galaktionov, and J. R. King, *Unstable sixth-order thin film equation.* II. *Global similarity patterns*, Nonlinearity, 20 (2007), pp. 1843–1881.

[20] J. D. Evans, V. A. Galaktionov, and J. F. Williams, *Blow-up and global asymptotics of the limit unstable Cahn–Hilliard equation*, SIAM J. Math. Anal., 38 (2006), pp. 64–102.

[21] R. Ferreira and F. Bernis, *Source-type solutions to thin-film equations in higher dimensions*, European J. Appl. Math., 8 (1997), pp. 507–524.

[22] V. A. Galaktionov, *Critical global asymptotics in higher-order semilinear parabolic equations*, Int. J. Math. Math. Sci., 60 (2003), pp. 3809–3825.

[23] V. A. Galaktionov and P. J. Harwin, *On evolution completeness of nonlinear eigenfunctions for the porous medium equation in the whole space*, Adv. Differential Equations, 10 (2005), pp. 635–674.

[24] V. A. Galaktionov and S. I. Pohozaev, *Blow-up and critical exponents for parabolic equations with non-divergent operators: Dual porous medium and thin film operators*, J. Evol. Equ., 6 (2006), pp. 45–69.

[25] V. A. Galaktionov and S. R. Svirshchevskii, *Exact Solution and Invariant Subspaces of Nonlinear Partial Differential Equations in Mechanics and Physics*, Chapman & Hall/CRC, Boca Raton, FL, 2007.

[26] V. A. Galaktionov and J. L. Vázquez, *A Stability Technique for Evolution Partial Differential Equations. A Dynamical Systems Approach*, Progr. Nonlinear Differential Equations Appl. 56, Birkhäuser Boston, Boston, 2004.

[27] L. Giacomelli and F. Otto, *Groplet spreading: Intermediate scaling law by PDE methods*, Comm. Pure Appl. Math., 55 (2002), pp. 217–254.

[28] L. V. Govor, J. Parisi, G. H. Bauer, and G. Reiter, *Instability and droplet formation in evaporating thin films of a binary solution*, Phys. Rev. E, 71 (2005), 051603.

[29] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.

[30] G. Grün, *Degenerate parabolic equations of fourth order and a plasticity model with non-local hardening*, Z. Anal. Anwendungen, 14 (1995), pp. 541–574.

[31] A. S. Kalashnikov, *Some problems of the qualitative theory of second-order nonlinear degenerate parabolic equations,* Russian Math. Surveys, 42 (1987), pp. 169–222.

[32] I. T. Kiguradze and T. Kusano, *Periodic solutions of nonautonomous ordinary differential equations of higher order*, Differential Equations, 35 (1999), pp. 71–77.

[33] M. A. Krasnosel'skii and P. P. Zabreiko, *Geometrical Methods of Nonlinear Analysis*, Springer-Verlag, Berlin, Tokyo, 1984.

[34] R. S. Laugesen and M. C. Pugh, *Energy levels of steady states for thin-film-type equations*, J. Differential Equations, 182 (2002), pp. 377–415.

[35] Z. Liu and Y. Mao, *Existence theorems for periodic solutions of higher order nonlinear differential equations*, J. Math. Anal. Appl., 216 (1997), pp. 481–490.

[36] V. G. Maz'ja, *Sobolev Spaces*, Springer-Verlag, Berlin, Tokyo, 1985.

[37] M. A. Naimark, *Linear Differential Operators*, Frederick Ungar, New York, 1968.

[38] A. Oron, *Nonlinear dynamics of three-dimensional long-wave Marangoni instability in thin liquid films*, Phys. Fluids, 12 (2000), pp. 1633–1645.

[39] A. Oron, S. H. Davies, and S. G. Bankoff, *Long-scale evolution of thin liquids films*, Rev. Modern Phys., 69 (1997), pp. 931–980.

[40] A. Oron and O. Gottlied, *Nonlinear dynamics of temporally excited falling liquid films*, Phys. Fluids, 14 (2002), pp. 2622–2636.

[41] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.

[42] N. F. Smyth and J. M. Hill, *High-order nonlinear diffusion*, IMA J. Appl. Math., 40 (1988), pp. 73–86.

[43] M. A. Vainberg and V. A. Trenogin, *Theory of Branching of Solutions of Non-linear Equations*, Noordhoff, Leiden, 1974.

[44] J. R. Ward, *Asymptotic conditions for periodic solutions of ordinary differential equations*, Proc. Amer. Math. Soc., 81 (1981), pp. 415–420.

[45] T. P. Witelski, A. J. Bernoff, and A. L. Bertozzi, *Blow-up and dissipation in a critical-case unstable thin film equation*, European J. Appl. Math., 15 (2004), pp. 223–256.

SIAM J. Appl. Math.
Vol. 69, No. 5, pp. 1359–1392

BURSTING OSCILLATIONS INDUCED BY SMALL NOISE*

PAWEL HITCZENKO† AND GEORGI S. MEDVEDEV†

Abstract. We consider a model of a square-wave bursting neuron residing in the regime of tonic spiking. Upon introduction of small stochastic forcing, the model generates irregular bursting. The statistical properties of the emergent bursting patterns are studied in the present work. In particular, we identify two principal statistical regimes associated with the noise-induced bursting. In the first case, type I, bursting oscillations are created mainly due to the fluctuations in the fast subsystem. In the alternative scenario, type II bursting, the random perturbations in the slow dynamics play a dominant role. We propose two classes of randomly perturbed slow-fast systems that realize type I and type II scenarios. For these models, we derive the Poincaré maps. The analysis of the linearized Poincaré maps of the randomly perturbed systems explains the distributions of the number of spikes within one burst and reveals their dependence on the small and control parameters present in the models. The mathematical analysis of the model problems is complemented by the numerical experiments with a generic Hodgkin–Huxley-type model of a bursting neuron.

Key words. neuronal dynamics, bursting, Hodgkin–Huxley model, slow-fast system, noise, Poincaré map

AMS subject classifications. 60H10, 34E10, 92C20

DOI. 10.1137/070711803

1. Introduction.

Differential equation models of excitable cells often include small random terms to reflect the unresolved or poorly understood aspects of the problem or to account for intrinsically stochastic factors [1, 8, 9, 10, 15, 16, 32, 41, 39, 43, 46]. In addition, many neuronal models also exhibit multistability [38, 26]. In systems with multiple stable states, noise may induce transitions between different attractors in the system dynamics, thus creating qualitatively new dynamical regimes that are not present in the deterministic system. In the present paper, we study this situation for a class of square-wave bursting models of excitable cell membranes. This class includes many conductance-based models of excitable cell membranes. Here we just mention the model of a pancreatic β-cell [6, 7], models of neurons in various central pattern generators such as those involved in insect locomotion [20], control of the heartbeat in a leech [25], and respiration in mammals [4, 5], to name a few. These models, as well as the underlying biological systems, exhibit characteristic bursting patterns of the voltage time series: clusters of fast spikes alternating with pronounced periods of quiescence (Figure 1a). For introduction to bursting, examples, and bibliography, we refer the reader to [26, 31, 37, 38, 44]. The dynamical patterns generated by the conductance-based models typically depend sensitively on parameters. For example, models of square-wave bursting neurons often exhibit both bursting and spiking behaviors for different values of parameters (see Figure 1a,b). In many relevant experiments, the transition from spiking to bursting is achieved by changing the injected current. In the present paper, we consider a model of a square-wave bursting neuron in the regime of tonic spiking (Figure 1b). We show that a small noise

*Received by the editors December 25, 2007; accepted for publication (in revised form) November 7, 2008; published electronically February 19, 2009. The work of the first author was partially supported by NSA grant MSPF-04G-054. The work of the second author was partially supported by NSF grant IOB 0417624.

http://www.siam.org/journals/siap/69-5/71180.html

†Department of Mathematics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 (phitczen@math.drexel.edu, medvedev@drexel.edu).

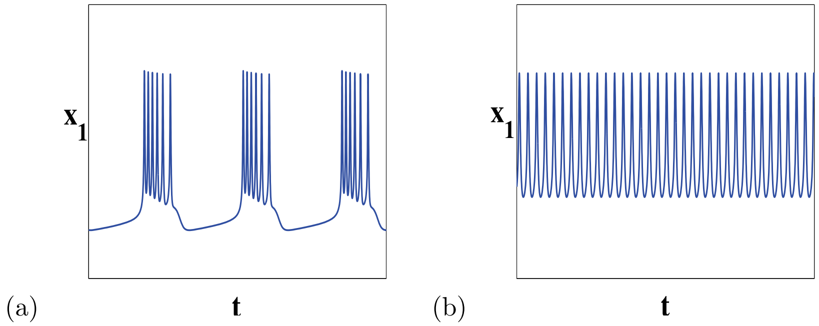


FIG. 1. *The dynamical patterns generated by a model of of a square-wave bursting neuron* (1.1) *and* (1.2)*:* (a) *periodic bursting and* (b) *tonic spiking.*

can transform spiking patterns into irregular (noise-induced) bursting patterns and describe two distinct mechanisms for generating noise-induced bursting. In the first scenario, bursting oscillations are triggered by the fluctuations in the fast subsystem. We refer to this mechanism as type I bursting. In contrast, the bursting dynamics in the type II scenario are driven by the random motion along the slow manifold. For each of these cases, we describe the statistical properties of the emergent bursting patterns and characterize them in terms of the small and control parameters present in the model.

Noise-induced phenomena have received considerable attention in the context of neuronal modeling (see, e.g., [1, 8, 9, 32, 39, 41, 43, 46]). A representative example is given by a two-dimensional (2D) excitable system perturbed by the white noise of small intensity [1]. In the presence of noise and under certain general conditions, a typical trajectory occasionally leaves the basin of attraction (BA) of the stable equilibrium and makes a large excursion in the phase plane of the deterministic system before returning to a small neighborhood of the stable fixed point (Figure 2a). This gives rise to irregular spiking (Figure 2b). The properties of the noise-induced spiking and stochastic resonance-type effects arising in the context of the perturbed FitzHugh–Nagumo model have been considered in [1, 8, 9, 10] (see also [3, 17, 18, 19] for the mathematical analysis of more general classes of related phenomena in randomly perturbed slow-fast systems). In the present paper, we study a related mechanism for irregular bursting. Specifically, we consider a class of models of square-wave bursting neurons:

$$\dot{x} = f(x, y), \tag{1.1}$$

$$\dot{y} = \epsilon g(x, y), \quad x = (x^1, x^2)^T \in \mathbb{R}^2,\ y \in \mathbb{R}^1, \tag{1.2}$$

where f and g are smooth functions and $0 < \epsilon \ll 1$ is a small parameter. We refer to (1.1), where y is treated as a parameter, as a fast subsystem. It is formally obtained from (1.1) and (1.2) by setting $\epsilon = 0$. We assume that the fast subsystem has a family of stable limit cycles and of stable equilibria for y in a certain interval $y \in (y_{sn}, y_{bp})$ (see Figure 3a). The additional assumptions on (1.1) and (1.2), which are explained in section 2, imply that for small $\epsilon > 0$, system (1.1) and (1.2) has a stable limit cycle, as shown in Figure 3c. In the presence of noise, a typical trajectory of the randomly perturbed system will occasionally leave the BA of the limit cycle of the deterministic system to make an excursion along the curve of equilibria of the degenerate system, E (see Figure 4a). Thus, in analogy to the 2D FitzHugh–Nagumo model (Figure 2a),

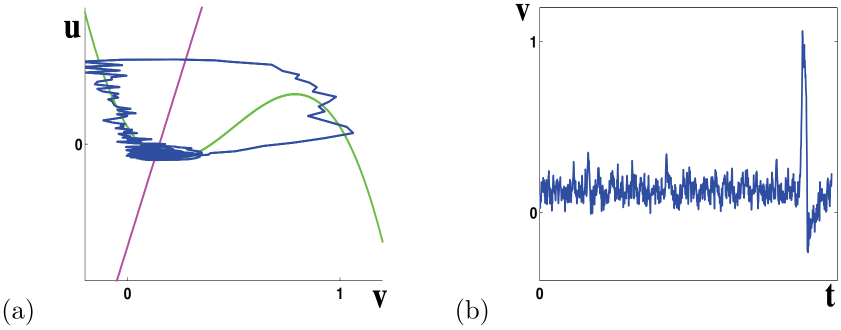


Fig. 2. (a) *A phase-plane trajectory of the randomly perturbed FitzHugh–Nagumo model in an excitable regime (see* [1] *for the model description and the parameter values).* (b) *The time series corresponding to the phase plot in* (a).

noise transforms spiking dynamics into irregular bursting. We refer to the latter as noise-induced bursting. In both examples above, irregular spiking (Figure 2a) or bursting patterns (Figure 4a,b) are created due to the escape of a trajectory of the randomly perturbed system from the BA of a stable fixed point in the case of spiking or of that of the stable limit cycle in the case of bursting. The statistics of the first exit times can then be related to the properties of the emergent firing patterns such as the frequency of spiking or the distribution of the number of spikes within one burst. Compared to the analysis of the irregular spiking in the randomly perturbed FitzHugh–Nagumo model (Figure 2), the analysis of the noise-induced bursting faces several additional challenges due to the fact that in the latter case one has to consider the exit problem for the trajectories near a stable limit cycle as opposed to those near a stable equilibrium in the former case. The structure of the BA of the limit cycle combined with the slow-fast character of the vector field determines the main features of the resultant bursting patterns. The description of the principal statistical regimes associated with the noise-induced bursting is the focus of the present paper.

There are general mathematical approaches for analyzing exit problems for stochastic processes generated by randomly perturbed differential equations such as (1.1) and (1.2): the Wentzell–Freidlin theory of large deviations [19] and the geometric theory for randomly perturbed slow-fast systems due to Berglund and Gentz [3]. In this paper we study the vector fields arising in the context of bursting. The specialized structure of this class of problems allows us to keep the analysis of the present paper self-contained and avoid using more technical methods, which are necessary for analyzing more general situations. Our analytical approach is based on the reduction of a randomly perturbed differential equation model to the Poincaré map and studying the exit problems for the trajectories of the discrete system. Using maps is quite natural in the context of bursting due to the intrinsic discreteness of bursting patterns imposed by the presence of spikes. Reductions to maps have been very useful for analyzing bursting dynamics in a variety of deterministic models [6, 33, 34, 35, 40]. As follows from the results of the present paper, the first return maps also provide a very convenient and visual representation for the mechanism underlying noise-induced bursting. In particular, we show that the distributions of spikes in one burst in many cases are effectively determined by one-dimensional (1D) linear randomly perturbed maps. We develop a set of probabilistic techniques for analyzing the dynamics of randomly perturbed 1D and 2D linear maps such as those arising in the analysis of bursting. The special structure of this class of problems, which is motivated by the applications to bursting, affords a more direct and simpler analysis than the treatment of more general classes of random linear maps found in the literature [29, 21, 28, 45].

The outline of the paper is as follows. In section 2, we formulate our assumptions on the deterministic system. We then present the preliminary numerical results, motivating our formulation of the randomly perturbed models at the end of this section. Specifically, we distinguish two types of the noise-induced bursting. *Type I* bursting is generated due to the fluctuations predominantly in the fast subsystem, while *type II* bursting is induced by variability mainly in the slow variable. Accordingly, we introduce two types of models that generate type I and type II bursting patterns. Section 3 develops a set of probabilistic techniques, which will be needed for the analysis of the first return maps for the randomly perturbed differential equation models. We first analyze a simple linear map with an attracting slope and small additive Gaussian perturbations in section 3.2. Due to the simple structure of the map, we obtain very explicit characterization of the first exit times for this problem. The analysis of this first relatively simple example provides the guidelines for the more complex cases dealt with in sections 3.3–3.5. Section 4 contains the definition and the construction of the Poincaré map for the type I randomly perturbed model introduced in section 2. The 2D Poincaré map is decomposed into two 1D maps for the fast and slow subsystems, which are constructed in sections 4.2 and 4.3, respectively. In section 4.4, we apply the results of section 3 to the linearization of the Poincaré map to derive the distributions of the first exit times. The latter are interpreted as the distributions of the number of spikes in one burst. In sections 4.5, we outline the modifications necessary to cover type II models. Since the analysis for type II models closely follows the lines of that for type I models, we omit most of the details. Finally, the numerical experiments in section 5 are designed to illustrate our theory.

2. The model.

In the present section, we introduce the model to be studied in the remainder of this paper. We start by formulating our assumptions on the deterministic model and then describe the random perturbation.

2.1. The deterministic model.

We consider slow-fast system (1.1) and (1.2) in $\mathbb{R}^3$ with one *slow* variable. The *fast* subsystem associated with (1.1) and (1.2) is obtained by sending $\epsilon \to 0$ in (1.2) and treating y as a parameter:

$$\dot{x} = f(x, y). \tag{2.1}$$

Under the variation of y, the fast subsystem has the bifurcation structure as shown schematically in Figure 3a. Specifically, we rely on the following assumptions:

(PO) There exists $y_{bp} \in \mathbb{R}$ such that for each $y < y_{bp}$, (2.1) has an exponentially stable limit cycle of period $\mathcal{T}(y)$:

$$L(y) = \{x = \phi(s, y) : \ 0 \leq s < \mathcal{T}(y)\}. \tag{2.2}$$

The family of the limit cycles, $L = \bigcup_{y<y_{bp}} L(y)$, forms a cylinder in $\mathbb{R}^3$ (Figure 3a).

(EQ) There is a branch of asymptotically stable equilibria of (2.1), $E = \{x = \psi(y) : y > y_{sn}\}$, which terminates at a saddle-node bifurcation at $y = y_{sn} < y_{bp}$ (Figure 3a).

(LS) For each $y \in \mathbb{R}$, the ω-limit set of almost all trajectories of (2.1) belongs to $L(y) \bigcup \{\psi(y)\}$.

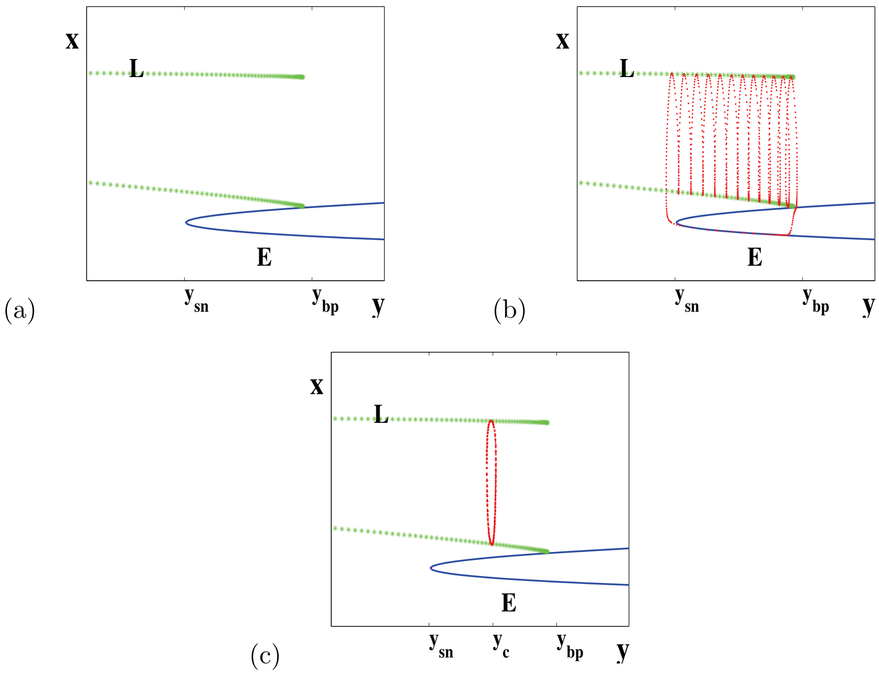


FIG. 3. (a) *The bifurcation diagram of the fast subsystem* (2.1). *L denotes a cylinder foliated by the stable periodic orbits. The lower branch of the parabolic curve E is composed of stable equilibria of the fast subsystem (see Figure* 6b *for the plot of a representative phase plane of the fast subsystem for $y \in (y_{sn}, y_{bp})$).* (b), (c) *Periodic trajectories of the full system* (1.1) *and* (1.2) *are superimposed on the bifurcation diagram of the fast subsystem. Assumptions* (SE) *and* (SB) *(see the text) result in a bursting limit cycle* (b), *while* (SS) *yields spiking* (c).

Remark 2.1. At $y = y_{bp}$, either L terminates or $L(y_{bp} + 0)$ loses stability. We do not specify the type of the bifurcation at $y = y_{bp}$. It may be, for instance, a homoclinic bifurcation as shown in Figure 3a, or a saddle-node bifurcation of limit cycles [22]. Having specified the assumptions on the bifurcation structure of the fast subsystem, we turn to the slow dynamics. The geometric theory for singularly perturbed systems implies the existence of the exponentially stable locally invariant manifolds E_ϵ and L_ϵ, which are $O(\epsilon)$ close to $E \bigcap \{(x, y) : y > y_{sn} + \delta\}$ and $L \bigcap \{(x, y) : y < y_{bp} - \delta\}$, respectively, for arbitrary fixed $\delta > 0$ and sufficiently small $\epsilon > 0$ [14, 27]. Manifolds E_ϵ and L_ϵ are called *slow manifolds*. For small $\epsilon > 0$, the dynamics of (1.1) and (1.2) on the slow manifolds is approximated by

$$L_\epsilon : \quad \dot{y} = \epsilon G(y), \quad y < y_{bp} - \delta, \tag{2.3}$$

$$E_\epsilon : \quad \dot{y} = \epsilon g(\psi(y), y), \quad y > y_{sn} + \delta, \tag{2.4}$$

where

$$G(y) = \frac{1}{\mathcal{T}(y)} \int_0^{\mathcal{T}(y)} g\left(\phi(s), y\right) ds. \tag{2.5}$$

We distinguish two types of the asymptotic behavior of solutions of (1.1) and (1.2): *bursting* and *spiking* (see Figure 1). The following conditions on the slow subsystem yield bursting.

For some $c > 0$ independent of ϵ,

(SE)

$$g(\psi(y), y) < -c \quad \text{for} \quad y > y_{sn}, \tag{2.6}$$

(SB)

$$G(y) > c \quad \text{for} \quad y < y_{bp}. \tag{2.7}$$

Under these assumptions, for sufficiently small $\epsilon > 0$ a typical trajectory of (1.1) and (1.2) consists of the alternating segments closely following L_ϵ and E_ϵ and fast transitions between them (see Figure 3b). For detailed discussions of the geometric construction of "bursting" periodic orbits, we refer the reader to [31, 37]. To obtain spiking, we substitute (SB) with

(SS) $G(y)$ has a unique simple zero at $y = y_c \in (y_{sn}, y_{bp})$:

$$G(y_c) = 0 \quad \text{and} \quad G'(y_c) < 0. \tag{2.8}$$

In this case, the asymptotic behavior of solutions follows from the following theorem due to Pontryagin and Rodygin.

Theorem 2.2 (see [36]). *If $\epsilon > 0$ is sufficiently small, system* (1.1) *and* (1.2) *has a unique exponentially stable limit cycle $L_\epsilon(y_c)$ of period $\mathcal{T}(y_c) + O(\epsilon)$ lying in an $O(\epsilon)$ neighborhood of $L(y_c)$, provided* (SS) *holds.*

Almost all trajectories of (1.1) and (1.2) are attracted by the limit cycle lying in an $O(\epsilon)$ neighborhood of $L(y_c)$. This mode of behavior is called spiking (see Figures 3c and 3b). In the remainder of this paper we assume (SS), in addition to (PO), (EQ), (LS), and (SE).

2.2. The randomly perturbed models.

In this subsection, we provide a heuristic description of the effects of the random perturbations on the dynamics of (1.1) and (1.2). To study these effects quantitatively, at the end of this section we propose two randomly perturbed models.

Suppose the trajectories of (1.1) and (1.2) experience weak stochastic forcing, such that the perturbed trajectories represent well-defined stochastic processes and are close to the trajectories of (1.1) and (1.2) on finite intervals of time. Since the trajectories of the unperturbed system remain in a small neighborhood of $L(y_c)$ (possibly after short transients), we expect that in the presence of noise the trajectories will occasionally leave the BA of $L(y_c)$ and after making a brief excursion along E will return back to the vicinity of $L(y_c)$. Therefore, under random perturbation the system can exhibit bursting dynamics, while the underlying deterministic system is in the spiking regime. We refer to this mode of behavior as *noise-induced* bursting. Our goal is to describe typical statistical regimes associated with the noise-induced bursting and to relate them to the structure of (1.1) and (1.2) and to the properties of the stochastic forcing. To illustrate the implications of the structure of the deterministic vector field for the bursting patterns that it produces under random perturbations, we refer to the following numerical examples. Note that the BA of $L(y_c)$ naturally extends along the cylinder of periodic orbits L (Figure 3c). The escape from the BA of $L(y_c)$ can be dominated by the fluctuations along L or by those in the transverse plane. These two possibilities are shown in Figure 4. The trajectory shown in Figure 4a spends most of the time near $L(y_c)$ and leaves its BA due to the fluctuations in the fast subsystem. We refer to this scenario as *type I* escape. Alternatively, the trajectory shown in Figure 4b travels a good deal along L before the escape and exits

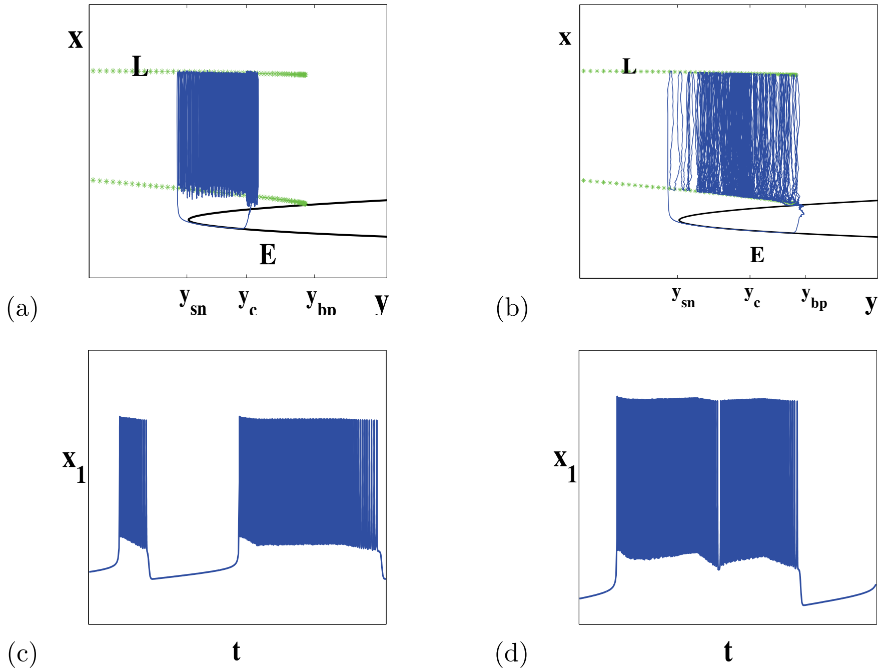


Fig. 4. *Noise-induced bursting.* (a) *A trajectory of the randomly perturbed system is shown in the phase space of the frozen system* (1.1), (1.2) *with* $\epsilon = 0$. *The trajectory leaves the basin of* $L(y_c)$ *mainly due to the fluctuations in the fast plane. This is characteristic to type I bursting. An alternative type II scenario is shown in plot* (b), *where the fluctuations in the slow direction dominate in the mechanism of escape from the basin of the stable limit cycle. The trajectory in* (b) *samples a wide region of* L *and leaves a neighborhood of* L *near the right boundary,* $y \approx y_{bp}$, *while that in* (a) *remains near* $L(y_c)$ *most of the time and jumps down near* $y \approx y_c$. *The differences translate into the distinctive features of the generic time series of the bursting patterns generated via type I or type II mechanisms shown in plots* (c) *and* (d), *respectively. Note that the longer burst in* (c) *has a typical square-wave form (roughly determined by* $L(y_c)$*), while the burst shown in* (d) *exhibits more variability due to the drifting of the trajectory along* L.

from the BA near $y = y_{bp}$. This mechanism is dominated by the slow dynamics. We refer to this scenario as *type II* escape. These mechanisms of escape translate into distinct features of the resultant bursting patterns. First, note that since in type I and type II scenarios the transition from spiking to quiescence typically takes place at $y \approx y_c$ and $y \approx y_{bp}$, respectively, by (1.2) and (EQ), the corresponding interburst intervals are approximately equal to

$$IBI \approx \epsilon^{-1} \int_{\hat{y}}^{y_{sn}} \frac{dy}{g\left(\psi(y), y\right)}, \quad \text{where} \quad \begin{cases} \hat{y} = y_c, & \text{type I,} \\ \hat{y} = y_{bp}, & \text{type II.} \end{cases}$$

In addition, we expect that the interspike intervals (ISIs) within one burst in type I scenarios are localized about $\mathcal{T}(y_c)$, since the trajectory of the randomly perturbed system in the active phase of bursting spends most of the time near $L(y_c)$. In type II bursting patterns, ISIs are expected to have more variability, since the trajectories sample a wider range of ISIs during their excursions along L. Perhaps, a more pronounced distinction between these two types of bursting patterns lies in the degree of the variability of the spikes in one burst. Most of the spikes forming a burst in type I pattern are generated by (2.1) with $y \approx y_c$ and, therefore, are similar in shape

(Figure 4c). In contrast, spikes in type II scenarios are subject to more variability and the bursting patterns typically have a ragged shape (Figure 4d).

To study type I and type II noise-induced bursting patterns it is convenient to consider two types of models. The *type I model* incorporates random forcing in the fast subsystem:

$$\dot{x}_t = f\left(x_t, y_t\right) + \sigma p \dot{w}_t, \tag{2.9}$$

$$\dot{y}_t = \epsilon g(x_t, y_t), \tag{2.10}$$

while in the *type II model* the slow subsystem is forced:

$$\dot{x}_t = f\left(x_t, y_t\right), \tag{2.11}$$

$$\dot{y}_t = \epsilon\left(g(x_t, y_t) + \sigma q \dot{w}_t\right). \tag{2.12}$$

Here, $0 < \sigma \ll 1$, $p(x,y) = \left(p^1(x,y), p^2(x,y)\right)^T$ and $q(x,y)$ are differentiable functions; $\dot{w}_t$ stands for the white noise, i.e., a generalized derivative of the Wiener process.

3. The randomly perturbed maps.

In this section, we develop probabilistic tools needed for the analysis of randomly perturbed systems (2.9)–(2.12). The number of spikes in one burst is a natural random variable associated with the noise-induced bursting. It is commonly used in the experimental studies of bursting, and we shall adopt it for characterizing irregular bursting patterns in this work. In section 4, we will show that the number of spikes in one burst is represented by a stopping time (more precisely, the level exceedance time) of a discrete random process, the Poincaré map of the randomly perturbed system (2.9)–(2.12). In preparation for the analysis of the linearized Poincaré map in section 4, in the present section we study certain stochastic linear difference equations. Equations of this form have been considered in the literature before. The study was initiated by Kesten [29], who considered the multidimensional case (in which the coefficients of the stochastic equations are random matrices). Subsequent work focused mostly on the 1D case. We refer the reader to the papers [21, 45], which contain representative results, examples of applications, and further references. There is also a review paper [12], unfortunately not easily accessible. The convergence properties of the solutions that we will need could be deduced from a general theory of stochastic difference equations. However, the results in the literature are often stated in the most general form and some of the proofs are rather involved. We will be dealing with special cases that are much easier to justify. For this reason, and also to keep the paper self-contained, we will include the proofs of the needed results.

3.1. Geometric random variables.

We begin by recalling the necessary properties of geometric random variables (RVs). Recall that Y is a geometric RV with parameter p, $0 < p < 1$, if

$$\mathbb{P}\left(Y = k\right) = p(1-p)^{k-1}, \quad k \geq 1. \tag{3.1}$$

We refer the reader to [28, Chapter 5] for the review of the properties of geometric distributions and their applications. In particular, the following characterization of geometric RVs is classical.

LEMMA 3.1. *Let Y be an RV with values in the set of positive integers. Y is a geometric with parameter p, $0 < p < 1$, iff*

$$\mathbb{P}(Y = n) = p\mathbb{P}(Y \geq n), \quad n \geq 1. \tag{3.2}$$

Lemma 3.1 motivates the following definition.

Definition 3.2. *Let Y be a random variable with values in the set of positive integers and let $0 < p < 1$. We say that Y is asymptotically geometric with parameter p if*

$$\lim_{n\to\infty} \frac{\mathbb{P}\,(Y = n)}{\mathbb{P}(Y \geq n)} = p. \tag{3.3}$$

3.2. The randomly perturbed map: Additive perturbation. Consider

$$Y_n = \lambda Y_{n-1} + \varsigma r_n, \quad n \geq 1, \tag{3.4}$$

where $r_1, r_2, \dots$ are independent identically distributed (IID) copies of the standard normal RV, and Y_0 is a real number. We will use $N(\mu, \eta^2)$ notation for a normal RV with mean μ, variance η^2, and probability density function given by

$$\frac{1}{\sqrt{2\pi}\eta} \exp\left\{-\frac{(x-\mu)^2}{2\eta^2}\right\}, \quad -\infty < x < \infty.$$

We will also let Z denote a generic $N(0,1)$ RV and write

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

for its distribution function. For a given $h > 0$, let

$$\tau = \inf\{k \geq 1 : \ Y_k > h\}.$$

Theorem 3.3. *Let*

$$\varepsilon \in (0,1), \quad \lambda = 1 - \varepsilon, \quad \beta^2 = \frac{\varsigma^2}{\varepsilon(2-\varepsilon)}, \quad \textit{and} \quad h - Y_0 > 0. \tag{3.5}$$

Then for sufficiently small $\varsigma > 0$, τ is asymptotically geometric RV with parameter

$$p = \frac{1}{\sqrt{2\pi}} \frac{\beta}{h\Phi(h/\beta)} \exp\left\{-\frac{h^2}{2\beta^2}\right\} \left(1 + O\left(\frac{\varsigma}{\varepsilon}\right)^2\right). \tag{3.6}$$

We precede the proof of the theorem with the following auxiliary.

Lemma 3.4. *For $n \geq 1$, Y_n is a normal RV with*

$$\mathbb{E}\ Y_n = \lambda^n Y_0 \quad \textit{and} \quad \operatorname{var} Y_n = \frac{\varsigma^2\left(1 - \lambda^{2n}\right)}{1 - \lambda^2} =: \beta_n^2. \tag{3.7}$$

In particular,

$$Y_n \xrightarrow{d} Y \overset{d}{=} N(0, \beta^2),$$

where $\xrightarrow{d}$ (and $\overset{d}{=}$) denote the convergence (equality) in distribution.

Proof of Lemma 3.4. The statements in (3.7) are verified by a straightforward calculation. The rest follows, because $\mathbb{E}\ Y_n \to 0$ and $\beta_n \to \beta$.

Proof of Theorem 3.3. Let $Y_k^* = \max\{Y_j : \ 1 \leq j \leq k\}$, $k \geq 1$. Then

$$\begin{aligned}\mathbb{P}(\tau = n+1) &= \mathbb{P}(Y_{n+1} > h, Y_n^* \leq h) = \mathbb{P}(Y_{n+1} > h | Y_n^* \leq h)\mathbb{P}(Y_n^* \leq h)\\ &= \mathbb{P}(Y_{n+1} > h | Y_n \leq h, Y_{n-1} \leq h, \dots, Y_0 \leq h)\mathbb{P}(\tau \geq n+1)\\ &= \mathbb{P}(Y_{n+1} > h | Y_n \leq h)\mathbb{P}(\tau \geq n+1). \end{aligned} \tag{3.8}$$

In the last equality, we used the fact that $\{Y_n\}$ is a Markov process which is clear from (3.4). By (3.8),

$$p_n := \frac{\mathbb{P}(\tau = n+1)}{\mathbb{P}(\tau \geq n+1)} = \mathbb{P}\left(Y_{n+1} > h \,|Y_n \leq h\,\right) = \frac{\mathbb{P}\left(Y_{n+1} > h, Y_n \leq h\right)}{\mathbb{P}\left(Y_n \leq h\right)}. \tag{3.9}$$

In accordance with Definition 3.2, we need to show that $\{p_n\}$ converges and to estimate the limit. By Lemma 3.4,

$$\mathbb{P}\left(Y_n \leq h\right) \longrightarrow \Phi(h/\beta), \text{ as } n \to \infty.$$

Next, we turn to estimating the numerator in (3.9). We have

$$\begin{aligned} Q_n &:= \mathbb{P}\left(Y_{n+1} > h, Y_n \leq h\right) = \mathbb{P}\left(\lambda Y_n + \varsigma r_{n+1} > h, Y_n \leq h\right) \\ &\to \mathbb{P}\left(\lambda Y + \varsigma Z > h, Y \leq h\right) =: Q, \end{aligned}$$

where Z is standard normal, Y is $N(0,\beta^2)$, and both are independent. This follows from Lemma 3.4 and the fact that r_{n+1} is $N(0,1)$ and is independent of Y_n. Q is the probability that a 2D Gaussian vector is in the region $[h,\infty) \times (-\infty, h]$. There are several ways of estimating this probability. We take the following, elementary approach. Let $X = h - Y$ so that X is $N\left(h,\beta^2\right)$ and is independent of Z. Then

$$Q = \mathbb{P}\left(Z > \frac{\varepsilon}{\varsigma}h + \frac{1-\varepsilon}{\varsigma}X, X \geq 0\right) = \frac{1}{\sqrt{2\pi}\beta}\int_0^\infty \mathbb{P}\left(Z > \frac{\varepsilon h + (1-\varepsilon)s}{\varsigma}\right) e^{\frac{-(s-h)^2}{2\beta^2}}\, ds.$$

By the well-known asymptotics (see [13, Chapter VII, Lemma 2 and section 7, Problem 1])

$$\mathbb{P}(Z > u) = 1 - \Phi(u) = \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{u^2}{2}}}{u}\left(1 + O\left(\frac{1}{u^2}\right)\right), \quad u > 0. \tag{3.10}$$

Hence, for sufficiently small $\varsigma > 0$ ($\varsigma \ll \varepsilon$), we have

$$Q \approx \frac{1}{2\pi}\frac{\varsigma}{\beta}\int_0^\infty \frac{\exp\left\{-\frac{1}{2}\left(\frac{(\varepsilon h + (1-\varepsilon)s)^2}{\varsigma^2} + \frac{(s-h)^2}{\beta^2}\right)\right\}}{\varepsilon h + (1-\varepsilon)s}\, ds. \tag{3.11}$$

Since

$$\frac{(\varepsilon h + (1-\varepsilon)s)^2}{\varsigma^2} + \frac{(s-h)^2}{\beta^2} = \frac{(s-\varepsilon h)^2}{\varsigma^2} + \frac{h^2}{\beta^2},$$

we obtain

$$Q \approx \frac{\varsigma}{2\pi\beta}\exp\left\{-\frac{h^2}{2\beta^2}\right\}\int_0^\infty \frac{\exp\left\{-\frac{(s-h\varepsilon)^2}{2\varsigma^2}\right\}}{\varepsilon h + (1-\varepsilon)s}\, ds.$$

By Laplace's method [47], for sufficiently small $\varsigma > 0$ ($\varsigma \ll \varepsilon$), the last integral is asymptotic to

$$\frac{\sqrt{2\pi}}{(h\varepsilon + (1-\varepsilon)\varepsilon h)\sqrt{1/\varsigma^2}} = \frac{\sqrt{2\pi}\varsigma}{h\varepsilon(2-\varepsilon)}.$$

Hence,

$$Q \approx \frac{\varsigma}{2\pi\beta} \frac{\sqrt{2\pi}\varsigma}{h\varepsilon(2-\varepsilon)} \exp\left\{-\frac{h^2}{2\beta^2}\right\} = \frac{\beta}{\sqrt{2\pi}h} \exp\left\{-\frac{h^2}{2\beta^2}\right\}.$$

By the same reasoning the error term from (3.10) is of order

$$\exp\left\{-\frac{h^2}{2\beta^2}\right\} \times O\left(\frac{1}{\varepsilon}\left(\frac{\beta}{h}\right)^3\right),$$

which gives (3.6). □

3.3. The randomly perturbed map: Random slope. Consider a process

$$Y_n = \mu(1+\sigma r_{1,n})Y_{n-1} + \sigma r_{2,n}, \quad n \geq 1, \tag{3.12}$$

where $(r_{1,n}, r_{2,n})_{n=1}^{\infty}$ are IID copies of a 2D random vector (r_1, r_2). Here, we assume that (r_1, r_2) has bivariate normal distribution with mean vector 0 and covariance matrix $\Sigma_2 = [\sigma_{i,j}]$, where $\sigma_{i,j} = \mathrm{cov}(r_i, r_j)$, $1 \leq i, j \leq 2$. We assume that the entries $\sigma_{i,j}$ are of order 1 in a sense that they do not depend on other parameters. Recall that the probability density function of a multivariate normal random vector $(r_1, \ldots, r_d)$ with mean vector 0 and covariance matrix Σ is given by

$$\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}x^T\Sigma^{-1}x\right\}, \quad x = (x_1, \ldots, x_d)^T,$$

and we denote such vectors by $N(0, \Sigma)$.

For a given $h > 0$, let

$$\tau = \inf\{k \geq 1 : \ Y_k > h\}.$$

THEOREM 3.5. *Suppose that h and $\mu \in (0,1)$ are both of order 1 and $\sigma \ll 1$ so that the following condition holds:*

$$\gamma := \mu\mathbb{E}|1+\sigma r_1| < 1. \tag{3.13}$$

Then τ is asymptotically geometric RV with parameter

$$p = \frac{\sigma}{c\sqrt{2\pi}} e^{-\frac{c^2}{2\sigma^2}}\Big(1 + O(\sigma^2)\Big), \tag{3.14}$$

where a positive constant c depends on h, μ, and Σ_2, but not on σ.

As before, we first establish convergence of $\{Y_n\}$ and characterize the limit. Iteration of (3.12) yields

$$\begin{aligned} Y_n &= \mu(1+\sigma r_{1,n})Y_{n-1} + \sigma r_{2,n} = \mu(1+\sigma r_{1,n})\left(\mu(1+\sigma r_{1,n-1})Y_{n-2} + \sigma r_{2,n-1}\right) + \sigma r_{2,n} \\ &= \cdots = \mu^n Y_0 \prod_{j=1}^{n}(1+\sigma r_{1,j}) + \sigma \sum_{j=0}^{n-1} \mu^j r_{2,n-j} \prod_{\ell=n-j+1}^{n} (1+\sigma r_{1,\ell}), \end{aligned} \tag{3.15}$$

where as usual $\prod_{j=k}^{m}(\ *\ ) = 1$ if $k > m$.

Lemma 3.6.

$$Y_n \xrightarrow{d} Y \overset{d}{=} \sigma \sum_{j=0}^{\infty} \mu^j g_{2,j} \prod_{\ell=0}^{j-1} (1+\sigma g_{1,\ell}),\ n\to\infty, \tag{3.16}$$

where $(g_{1,j}, g_{2,j}),\ j = 0, 1, 2, \ldots,$ *are IID copies of a* 2D *random vector, which is equal in distribution to* (r_1, r_2)*.*

Proof of Lemma 3.6. First, we show that Y is well-defined as the series in (3.16) converges almost surely. To see this, note that the summands

$$g_{2,j} \prod_{\ell=0}^{j-1} (1+\sigma g_{1,\ell})$$

are martingale differences with respect to the natural filtration. By triangle inequality, independence, and (3.13),

$$\begin{aligned}\mathbb{E}\left|\sigma \sum_{j=0}^{m} \mu^j g_{2,j} \prod_{\ell=0}^{j-1}(1+\sigma g_{1,\ell})\right| &\le \sigma \mathbb{E}|g_2| \sum_{j=0}^{m} \mu^j \mathbb{E}\left|\prod_{\ell=0}^{j-1}(1+\sigma g_{1,\ell})\right| \\ &= \sigma\mathbb{E}|g_2| \sum_{j=0}^{m} \mu^j \left(\mathbb{E}|1+\sigma r_1|\right)^j = \frac{\sigma\mathbb{E}|g_2|}{1-\gamma}(1-\gamma^{(m+1)}) \le \frac{\sigma \mathbb{E}|g_2|}{1-\gamma}.\end{aligned}$$

Hence, the partial sums of the right-hand side of (3.16) form an L_1-bounded martingale which converges almost surely by the martingale convergence theorem (see, e.g., [42]). For every $n \ge 1$

$$\sigma \sum_{j=0}^{n-1} \mu^j r_{2,n-j} \prod_{\ell=n-j+1}^{n} (1+\sigma r_{1,\ell}) \overset{d}{=} \sigma \sum_{j=0}^{n-1} \mu^j g_{2,j} \prod_{\ell=0}^{j-1} (1+\sigma g_{1,\ell}).$$

Since the sequence on the right converges almost surely and the almost sure convergence implies convergence in distribution, we infer that the sequence on the left converges in distribution. To conclude that $Y_n \xrightarrow{d} Y$ it is enough to show that the first term on the right-hand side of (3.15) converges to 0 in probability. But that is clear since we have

$$\mathbb{E}\left|Y_0 \mu^n \prod_{j=1}^{n}(1+\sigma r_{1,j})\right| = |Y_0|\mu^n \prod_{j=1}^{n} \mathbb{E}|1+\sigma r_{1,j}| = |Y_0|\gamma^n.$$

Hence, by Markov inequality it goes to 0 in probability. □

Proof of Theorem 3.5. The proof follows along the lines of the proof of Theorem 3.3. The main complication in treating the present case is that we know less about the distribution of Y_n than before. Nonetheless, we will argue that for large n

$$p_n := \frac{\mathbb{P}(\tau = n)}{\mathbb{P}(\tau \ge n)} = \mathbb{P}(\mu(1+\sigma r_{1,n})Y_{n-1} + \sigma r_{2,n} > h | Y_{n-1} \le h) \tag{3.17}$$

is approximately constant. For this, we rewrite the right-hand side of (3.17) as

$$\frac{\mathbb{P}(\mu(1+\sigma r_{1,n})Y_{n-1} + \sigma r_{2,n} > h, Y_{n-1} \le h)}{\mathbb{P}(Y_{n-1} \le h)},$$

and since the denominator converges to $\mathbb{P}(Y \le h)$ we focus on the numerator. Let (r_1, r_2) be a generic vector distributed like $(r_{1,n}, r_{2,n})$ and independent of Y. Since for every $n \ge 1$, $(r_{1,n}, r_{2,n})$ is independent of Y_{n-1}, as $n \to \infty$, we have

$$(r_{1,n}, r_{2,n}, Y_{n-1}) \xrightarrow{d} (r_1, r_2, Y).$$

Thus,

$$\mathbb{P}(\mu(1+\sigma r_1)Y_{n-1} + \sigma r_2 > h, Y_{n-1} \le h) \longrightarrow \mathbb{P}(\mu(1+\sigma r_1)Y + \sigma r_2 > h, Y \le h),$$

which establishes the existence of $p = \lim_{n\to\infty} p_n$.

To estimate p, we first recall that (r_1, r_2) is bivariate normal iff every linear combination of r_1 and r_2 is a normal RV. Hence, conditionally on $Y = y$, $\sigma(\mu y r_1 + r_2)$ is $N(0, \sigma^2\sigma_y^2)$ RV, where

$$\sigma_y^2 = \sigma_{22}^2 + \mu^2 y^2 \sigma_{11}^2 + 2\mu y \sigma_{12}. \tag{3.18}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(\mu(1+\sigma r_1)Y + \sigma r_2 > h,\ Y \le h) = \mathbb{P}(\sigma(\mu Y r_1 + r_2) > h - \mu Y, Y \le h) \\
&\quad = \int_{-\infty}^{h} \mathbb{P}\left(Z > \frac{h-\mu y}{\sigma\sigma_y}\right) dF_Y(y) = \int_{-\infty}^{h} \left(1 - \Phi\left(\frac{h-\mu y}{\sigma\sigma_y}\right)\right) dF_Y(y) \\
&\quad = \left(1 - \Phi\left(\frac{h-\mu y_0}{\sigma\sigma_{y_0}}\right)\right) \mathbb{P}(Y \le h),
\end{aligned}$$

where $-\infty < y_0 < h$ by the mean value theorem. Hence,

$$p = \frac{\mathbb{P}(\mu(1+\sigma r_1)Y + \sigma r_2 > h,\ Y \le h)}{\mathbb{P}(Y \le h)} = 1 - \Phi\left(\frac{h-\mu y_0}{\sigma\sigma_{y_0}}\right).$$

Let $c := c(y_0)$, where

$$c(x) = c_{h,\mu,\Sigma_2}(x) := \frac{h-\mu x}{\sigma_x} = \frac{h-\mu x}{\sqrt{\mu^2\sigma_{11}^2 x^2 + 2\mu\sigma_{12}x + \sigma_{22}^2}}.$$

Then, by (3.10),

$$p = 1 - \Phi\left(\frac{c}{\sigma}\right) = \frac{\sigma}{c\sqrt{2\pi}} e^{-\frac{c^2}{2\sigma^2}} \left(1 + O\left(\frac{\sigma^2}{c^2}\right)\right).$$

Furthermore, by elementary analysis we see that

- $c(x)$ is increasing on $x \in (-\infty, x^*)$ and decreasing on $x \in (x^*, \infty)$, where
$$x^* = -\frac{\sigma_{11}^2 + h\sigma_{12}}{\mu(h\sigma_{22}^2 + \sigma_{12})};$$
- $c(-\infty) = \sigma_{11}^{-1}$, $c(h) = \frac{(1-\mu)h}{((\mu h\sigma_{11})^2 + 2\mu\sigma_{12}h + \sigma_{22})^{1/2}} = \frac{(1-\mu)h}{((\mu h\sigma_{11} + \sigma_{22})^2 - 2\mu h(\sigma_{11}\sigma_{22} - \sigma_{12}))^{1/2}}$, and $c(x^*)$ is given by a quite unwieldy expression that depends on h and Σ_2 but not on μ.

In particular, c is bounded away from 0 and ∞, provided μ and h are positive and $\mu < 1$. This proves (3.14). □

3.4. A 2D randomly perturbed map.

In this subsection we consider the following 2D model:

$$\xi_{n+1} = \mu\xi_n \left(1 + \sigma r_{1,n+1}\right) + \sigma r_{2,n+1}, \tag{3.19}$$

$$\eta_{n+1} = \lambda\eta_n + \epsilon\sigma r_{3,n+1} + \epsilon a_2 \xi_n, \tag{3.20}$$

where $(r_{1,n}, r_{2,n}, r_{3,n})$, $n \geq 1$, is a sequence of IID copies of (r_1, r_2, r_3) which, as follows from a discussion at the beginning of section 4.4, is assumed to be a trivariate normal random vector $N(0, \Sigma_3)$, with $\Sigma_3 = [\sigma_{i,j}]$, $1 \leq i, j \leq 3$, where $\sigma_{i,j} = \operatorname{cov}(r_i, r_j)$ do not depend on any parameters in (4.44) and (4.45). For positive $h_1, h_2 = O(1)$, we define

$$\tau_\xi = \inf_{k \geq 1}\{\xi_k > h_1\}, \qquad \tau_\eta = \inf_{k \geq 1}\{\eta_k > h_2\}.$$

We are interested in $\tau = \min\{\tau_\xi, \tau_\eta\}$. We know the distribution of τ_ξ from Theorem 3.5. As we will show below, under suitable conditions the distribution of τ is again asymptotically geometric. Moreover, if $\epsilon > 0$ is small, then τ_η has practically no effect on the distribution of τ.

In order to be more precise, let us define

$$A_n = \begin{bmatrix} \mu(1 + \sigma r_{1,n}) & 0 \\ \epsilon a_2 & \lambda \end{bmatrix}, \quad G_n = \begin{bmatrix} r_{2,n} \\ \epsilon r_{3,n} \end{bmatrix}, \quad \text{and} \quad \Theta_n = \begin{bmatrix} \xi_n \\ \eta_n \end{bmatrix}. \tag{3.21}$$

Then (3.19) and (3.20) are described by

$$\Theta_{n+1} = A_{n+1}\Theta_n + \sigma G_{n+1}, \quad n \geq 1. \tag{3.22}$$

Theorem 3.7. *Let $\mu, \sigma, \epsilon \in (0, 1)$ be such that μ is of order 1 and $\sigma \ll 1$ so that condition (3.13) holds. Assume $\epsilon \ll 1$ and set $\lambda = 1 - \epsilon$. Suppose further that h_1 and h_2 are of order 1. Then τ is an approximately geometric RV with parameter p satisfying*

$$p \approx \frac{\sigma}{c\sqrt{2\pi}} e^{-\frac{c^2}{2\sigma^2}}, \tag{3.23}$$

and where the constant c depends on h_1, μ, and Σ_3 but not on σ.

The following lemma shows that $\{\Theta_n\}$ converges in distribution and describes the limit.

Lemma 3.8.

$$\Theta_n \xrightarrow{d} X \overset{d}{=} \sigma \sum_{k=1}^{\infty} \left(\prod_{j=1}^{k-1} A_j\right) G_k, \; n \to \infty, \tag{3.24}$$

where A_n and $G_n, n = 1, 2, \ldots,$ are defined in (3.21). Furthermore, this random vector X satisfies the distributional equation

$$X \overset{d}{=} AX + \sigma G, \tag{3.25}$$

where

$$A = \begin{bmatrix} \mu(1 + \sigma r_1) & 0 \\ \epsilon a_2 & \lambda \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} r_2 \\ \epsilon r_3 \end{bmatrix}, \tag{3.26}$$

(r_1, r_2, r_3) is $N(0, \Sigma_3)$, and X on the right-hand side of (3.25) *is independent of* (A, G).

Proof of Lemma 3.8. Note first that each of the sequences (A_n) and (G_n) consists of IID random elements. By iterating (3.22), we obtain

$$\Theta_n = A_n(A_{n-1}\Theta_{n-2}+\sigma G_{n-1})+\sigma G_n = \cdots = \left(\prod_{k=0}^{n-1} A_{n-k}\right)\Theta_0+\sigma\sum_{k=1}^{n}\left(\prod_{j=0}^{n-k-1} A_{n-j}\right) G_k,$$

where, as usual, the product is set to be 1 if its index range is empty. We have

$$\prod_{k=0}^{n-1} A_{n-k} = \begin{bmatrix} \mu^n \prod_{k=1}^{n}(1+\sigma r_{1,k}) & 0 \\ T_n & \lambda^n \end{bmatrix},$$

where

$$T_n = \epsilon a_2 \sum_{j=1}^{n} \lambda^{n-j} \prod_{k=1}^{j-1}(\mu(1+\sigma r_{1,k})).$$

Set $\delta = \max\{\lambda, \mu\mathbb{E}|1+\sigma r_1|\}$ and note that by (3.13) $\delta < 1$. By triangle inequality and independence of $r_{1,k}$'s

$$\mathbb{E}|T_n| \le \epsilon a_2 \sum_{j=1}^{n} \lambda^{n-j}\mathbb{E}\left|\prod_{k=1}^{j-1}(\mu(1+\sigma r_{1,k}))\right| = \epsilon a_2 \sum_{j=1}^{n} \lambda^{n-j}\left(\mu\mathbb{E}|1+\sigma r_1|\right)^{j-1} \le \epsilon a_2 n\delta^{n-1}.$$

Similarly,

$$\mu^n\mathbb{E}\left|\prod_{k=1}^{n}(1+\sigma r_{1,k})\right| = \left(\mu\mathbb{E}|1+\sigma r_1|\right)^n \le \delta^n.$$

It follows that both components of $\big(\prod_{k=0}^{n-1} A_{n-k}\big)\Theta_0$ converge to 0 in probability, and thus this term is negligible.

Since the sequences (A_n) and (G_n) are IID, for every $n \ge 1$ we have

$$\sum_{k=1}^{n}\left(\prod_{j=0}^{n-k-1} A_{n-j}\right) G_k \overset{d}{=} \sum_{k=1}^{n}\left(\prod_{j=1}^{k-1} A_j\right) G_k.$$

By the same argument as above we verify that both components of the sequence of partial sums on the right-hand side are Cauchy in L_1. Hence, the components of the series

$$\sum_{k=1}^{\infty}\left(\prod_{j=1}^{k-1} A_j\right) G_k$$

converge in probability (and thus in distribution). Therefore, the sequence (Θ_n) defined by (3.22) converges in distribution to a random vector X defined in (3.24). Furthermore, X satisfies the distributional equation (3.25). □

Proof of Theorem 3.7. For $h = (h_1, h_2)$ set $B_h := (-\infty, h_1] \times (-\infty, h_2]$. Then

$$\{\tau = n\} = \{\Theta_j \in B_h,\ j < n,\ \Theta_n \notin B_h\},$$

so that

$$\begin{aligned}\mathbb{P}(\tau = n) &= \mathbb{P}(\Theta_n \notin B_h | \Theta_j \in B_h,\ j < n)\mathbb{P}(\Theta_j \in B_h,\ j < n)\\ &= \mathbb{P}(A_n\Theta_{n-1} + \sigma G_n \notin B_h | \Theta_{n-1} \in B_h)\mathbb{P}(\tau \geq n).\end{aligned}$$

Since Θ_n converge in distribution to X, we have

$$\begin{aligned}p_n := \mathbb{P}(A_n\Theta_{n-1} + \sigma G_n \notin B_h | \Theta_{n-1} \in B_h) &= \frac{\mathbb{P}(A_n\Theta_{n-1} + \sigma G_n \notin B_h, \Theta_{n-1} \in B_h)}{\mathbb{P}(\Theta_{n-1} \in B_h)}\\ \longrightarrow p := \frac{\mathbb{P}(AX + \sigma G \notin B_h, X \in B_h)}{\mathbb{P}(X \in B_h)}, &\quad \text{as} \quad n \to \infty. \qquad (3.27)\end{aligned}$$

It follows from (3.24) that X is symmetric, so since both h_1 and h_2 are positive the denominator is at least $1/2$ and does not affect the asymptotics.

To handle the numerator, using (3.26), denoting the components of X by X_1 and X_2, and using the notation adopted in (3.18), we see that it is equal to

$$\begin{aligned}&\mathbb{P}((\mu(1+\sigma r_1)X_1 + \sigma r_2, \epsilon a_2 X_1 + \lambda X_2 + \epsilon\sigma r_3) \notin B_h, (X_1, X_2) \in B_h)\\ &\quad = \mathbb{P}(\mu(1+\sigma r_1)X_1 + \sigma r_2 > h_1, (X_1, X_2) \in B_h)\\ &\qquad + \mathbb{P}(\epsilon a_2 X_1 + \lambda X_2 + \epsilon\sigma r_3 > h_2, (X_1, X_2) \in B_h)\\ &\qquad - \mathbb{P}(\mu(1+\sigma r_1)X_1 + \sigma r_2 > h_1, \epsilon a_2 X_1 + \lambda X_2 + \epsilon\sigma r_3 > h_2, (X_1, X_2) \in B_h)\\ &\quad = \mathbb{P}\left(\frac{\mu X_1 r_1 + r_2}{\sigma_{X_1}} > \frac{h_1 - \mu X_1}{\sigma\sigma_{X_1}}, (X_1, X_2) \in B_h\right)\\ &\qquad + \mathbb{P}\left(r_3 > \frac{h_2 - \epsilon a_2 X_1 - \lambda X_2}{\epsilon\sigma}, (X_1, X_2) \in B_h\right)\\ &\qquad - \mathbb{P}\left(\frac{\mu X_1 r_1 + r_2}{\sigma_{X_1}} > \frac{h_1 - \mu X_1}{\sigma\sigma_{X_1}}, r_3 > \frac{h_2 - \epsilon a_2 X_1 - \lambda X_2}{\epsilon\sigma}, (X_1, X_2) \in B_h\right). \qquad (3.28)\end{aligned}$$

Conditionally on $(X_1, X_2) = (x_1, x_2)$,

$$Z_1 := \frac{\mu x_1 r_1 + r_2}{\sigma_{x_1}} \quad \text{and} \quad Z_2 := \frac{r_3}{\sigma_{33}}$$

are $N(0,1)$ RVs. Hence by letting $F_X(x_1, x_2)$ denote the distribution function of (X_1, X_2), we see that the first of the last three probabilities is

$$\int_{-\infty}^{h_2}\int_{-\infty}^{h_1}\left(1 - \Phi\left(\frac{h_1 - \mu x_1}{\sigma\sigma_{x_1}}\right)\right) dF_X(x_1, x_2). \qquad (3.29)$$

Likewise, for the second of these probabilities we get

$$\int_{-\infty}^{h_2}\int_{-\infty}^{h_1}\left(1 - \Phi\left(\frac{h_2 - \epsilon a_2 x_1 - \lambda x_2}{\epsilon\sigma\sigma_{33}}\right)\right) dF_X(x_1, x_2). \qquad (3.30)$$

We now note that if ϵ is of a smaller order than all other parameters (except possibly σ), then (3.10) implies that (3.30) (and hence also (3.28)) are negligible when compared to (3.29). To analyze the behavior of (3.29) as a function of its parameters, note that by the mean value theorem the quantity in (3.29) is equal to

$$\left(1 - \Phi\left(\frac{h_1 - \mu x_0}{\sigma\sigma_{x_0}}\right)\right)\int_{-\infty}^{h_2}\int_{-\infty}^{h_1} dF_X(x_1, x_2) = \left(1 - \Phi\left(\frac{h_1 - \mu x_0}{\sigma\sigma_{x_0}}\right)\right)\mathbb{P}(X \in B_h)$$

for some $-\infty < x_0 < h$. Substituting this into (3.27) (and neglecting the terms that depend on ϵ) we see that

$$p = \frac{\mathbb{P}(AX + \sigma G \notin B_h, X \in B_h)}{\mathbb{P}(X \in B_h)} \sim 1 - \Phi\left(\frac{h_1 - \mu x_0}{\sigma \sigma_{x_0}}\right).$$

If both $0 < \mu < 1$ and h_1 are of order 1, we are in the same situation as with (3.14). This shows (3.23). $\square$

3.5. Diffusive escape. The exit problems for the stochastic difference equations analyzed in the previous subsections all feature the geometric escape mechanism. In the simplest case when the evolution is given by (3.4), the geometric distribution characterizes the statistics of the times of exit of the trajectories of (3.4) from a certain neighborhood of the attracting fixed point. In this subsection, we study another statistical regime associated with the exit problem for (3.4) that is important in applications: the diffusive regime. The role of the diffusive regime in characterizing the statistics of the exit times for the trajectories of (3.4) is twofold. First, the geometric distribution approximates the distribution of the exit times only for sufficiently large times, i.e., for large n. In this subsection, we show that in the intermediate range of n, i.e., when n is neither too large nor too small, Y_n's are approximated by the sums of the IID RVs, and therefore the level exceedance times are distributed as those for random walks. We refer to this situation as the diffusive regime. Second, we recall that to justify the geometric distribution in the proof of Theorem 3.3, we implicitly assumed that the rate of attraction of the fixed point is stronger than the noise intensity. Specifically, it is easy to see from the proof of Theorem 3.3 that ς is required to be $o(\epsilon)$, $\epsilon = 1 - \lambda$. The analysis in this subsection does not use this assumption. We show that when the noise is stronger than the attraction of the fixed point (though both are sufficiently small), the mechanism of escape of the trajectories from the BA of the fixed point changes from geometric to diffusive. Therefore, we conclude this section by pointing out some features intrinsic to the diffusive escape. Specifically, we consider (3.4), for which we define, as before,

$$\tau = \inf\{k \geq 1 : \; Y_k > h\} \tag{3.31}$$

for given $h > 0$. In contrast to the case considered in section 3.2, here we assume

$$\varepsilon = O(\varsigma^{\alpha}), \quad \alpha > 0. \tag{3.32}$$

In Theorem 3.9 below, we show that in the present situation in the intermediate range of n, $Y_n's$ behave as sums of IID normal RVs. The behavior of the latter is well known (cf. Lemma 3.11).

Recall that $\Phi(x)$ stands for the distribution function of an $N(0,1)$ RV and denote

$$\Psi_a(x) = 2\left(1 - \Phi\left(\frac{a}{\sqrt{x}}\right)\right), \quad a > 0. \tag{3.33}$$

Note that $\Psi_a(x)$ is a probability distribution function on $\mathbb{R}^+$ (see Figure 5).

THEOREM 3.9. *Let the evolution of* Y_n, $n = 0, 1, 2, \ldots$, *be given by* (3.4). *Suppose that* $\lambda = 1 - \varepsilon$ *with* $\varepsilon = O(\varsigma^{\alpha})$, $\alpha > 0$. *Then for arbitrary positive* β_1 *and* β_2 *such that* $\beta_1 + \beta_2 < 2\alpha/3$, *for sufficiently small* $\varsigma > 0$,

$$\mathbb{P}(\tau \leq n) = \Psi_a(n)\Big(1 + o(1)\Big), \quad a = \frac{h}{\varsigma}, \tag{3.34}$$

in the range $\varsigma^{-\beta_1} \ll n \ll \varsigma^{\frac{-2\alpha}{3} + \beta_2}$.

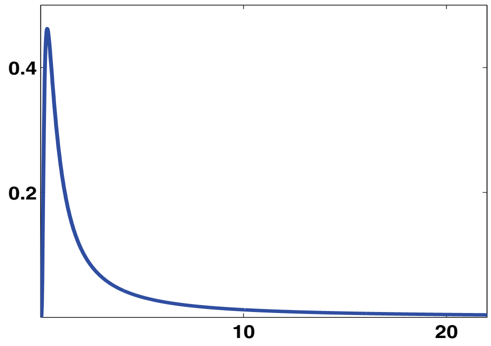


FIG. 5. *Probability density function corresponding to the distribution $\Psi_a(y), a = 1$. With a suitable $a > 0$, $\Psi_a(y)$ approximates the distribution of the exit times in the diffusive escape.*

Remark 3.10. Since $\beta_{1,2} > 0$ are arbitrary, $\Psi_a(n)$ practically approximates $\mathbb{P}(\tau \leq n)$ in the range $1 \ll n \ll \varepsilon^{-2/3}$.

We will need the following auxiliary lemma [11, Theorem 2.2, Chapter III]. It may be viewed as a quantified version of a reflection principle for random walk (see, e.g., [42, sections 5.3, 5.4]).

LEMMA 3.11. *Let $X_1, X_2, \dots$ be a sequence of independent, symmetric RVs and set*

$$S_k = \sum_{j=1}^{k} X_j \quad \textit{and} \quad S_k^* = \max_{1\leq j\leq k} S_j, \quad j \geq 1.$$

Then for any $t, u > 0$ the following inequalities hold:

$$2\mathbb{P}(S_n \geq t + 2u) - 2\sum_{k=1}^{n} \mathbb{P}(X_k \geq u) \leq \mathbb{P}(S_n^* \geq t) \leq 2\mathbb{P}(S_n \geq t). \tag{3.35}$$

Remark 3.12. As was noticed by Kwapień, a bit stronger version of the first inequality in (3.35) follows from a slight modification of the proof of Proposition 1.3.1 in [30].

Proof of Theorem 3.9. Without loss of generality, we assume that $Y_0 = 0$ (otherwise, apply the same argument to $Y_k - Y_0$). Note that the distributions of τ and Y_k^* are linked by the following relation:

$$\mathbb{P}(\tau \leq n) = \mathbb{P}(Y_n^* \geq h).$$

Unwinding (3.4) and using $Y_0 = 0$ gives

$$Y_k = \varsigma(\lambda^{k-1} r_1 + \lambda^{k-2} r_2 + \cdots + \lambda r_{k-1} + r_k),$$

which we write as $S_k + W_k$, where

$$S_k := \varsigma \sum_{j=1}^{k} r_j, \qquad W_k := \varsigma \sum_{j=1}^{k-1} r_j(\lambda^{k-j} - 1). \tag{3.36}$$

We will first show that the main contribution to Y_n^* is from the S_n^*. First, by subadditivity of maxima, for any $0 < h_1 < h$,

$$\begin{aligned}\mathbb{P}(Y_n^* \geq h) \leq \mathbb{P}(S_n^* + W_n^* \geq h) &\leq \mathbb{P}(S_n^* \geq h - h_1) + \mathbb{P}(W_n^* \geq h_1) \\ &\leq \mathbb{P}(S_n^* \leq h - h_1) + \mathbb{P}\left(|W_n|^* \geq h_1\right).\end{aligned} \tag{3.37}$$

Further, $Y_k \geq S_k - |W_k|$ so that

$$\begin{aligned}\mathbb{P}(S_n^* \geq h + h_1) &\leq \mathbb{P}(S_n^* \geq h + h_1, |W_n|^* < h_1) + \mathbb{P}(|W_n|^* \geq h_1) \\ &\leq \mathbb{P}(Y_n^* \geq h) + \mathbb{P}(|W_n|^* \geq h_1),\end{aligned}$$

which, when combined with (3.37), means that

$$\mathbb{P}(S_n^* \geq h + h_1) - \mathbb{P}(|W_n|^* \geq h_1) \leq \mathbb{P}(Y_n^* \geq h) \leq \mathbb{P}(S_n^* \geq h - h_1) + \mathbb{P}(|W_n|^* \geq h_1). \tag{3.38}$$

First, we estimate $\mathbb{P}(|W_n|^* \geq h_1)$ in (3.38). To this end, we use $1-\lambda^j = 1-(1-\varepsilon)^j \leq j\varepsilon$ to obtain

$$\operatorname{var}(W_n) = \varsigma^2 \sum_{j=1}^{n-1} (1-\lambda^j)^2 \leq \varsigma^2 \varepsilon^2 \frac{n^3}{3} = \varsigma^2 n \frac{\varepsilon^2 n^2}{3}.$$

Consequently, by (3.35) and (3.36), we have

$$\begin{aligned}\mathbb{P}(|W_n|^* \geq h_1) \leq 2\mathbb{P}(|W_n| \geq h_1) \leq 4\mathbb{P}(W_n \geq h_1) &= 4\mathbb{P}\left(Z \geq \frac{h_1}{\sqrt{\operatorname{var}(W_n)}}\right) \\ &\leq 4\mathbb{P}\left(Z \geq \frac{h_1}{\varsigma\sqrt{n}} \cdot \frac{\sqrt{3}}{\varepsilon n}\right).\end{aligned}$$

Next, we turn to estimating the probabilities involving S_n^* in (3.38). By the second inequality in (3.35), for every $u > 0$, we have

$$\mathbb{P}(S_n^* \geq h - h_1) \leq 2\mathbb{P}(S_n \geq h - h_1) = 2\mathbb{P}\left(Z \geq \frac{h - h_1}{\varsigma\sqrt{n}}\right), \tag{3.39}$$

while the first one yields

$$\begin{aligned}\mathbb{P}(S_n^* \geq h + h_1) &\geq 2\mathbb{P}\left(S_n \geq h + h_1 + 2u\right) - 2\sum_{k=1}^{n} \mathbb{P}(\varsigma r_k \geq u) \\ &= 2\mathbb{P}\left(Z \geq \frac{h + h_1 + 2u}{\varsigma\sqrt{n}}\right) - 2n\mathbb{P}\left(Z \geq \frac{u}{\varsigma}\right).\end{aligned} \tag{3.40}$$

The combination of (3.38), (3.39), and (3.40) yields

$$\mathbb{P}(Y_n^* \geq h) \geq 2\mathbb{P}\left(Z \geq \frac{h + h_1 + 2u}{\varsigma\sqrt{n}}\right) - 2n\mathbb{P}\left(Z \geq \frac{u}{\varsigma}\right) - 4\mathbb{P}\left(Z \geq \frac{h_1}{\varsigma\sqrt{n}} \cdot \frac{\sqrt{3}}{\varepsilon n}\right), \tag{3.41}$$

$$\mathbb{P}(Y_n^* \geq h) \leq 2\mathbb{P}\left(Z \geq \frac{h - h_1}{\varsigma\sqrt{n}}\right) + 4\mathbb{P}\left(Z \geq \frac{h_1}{\varsigma\sqrt{n}} \cdot \frac{\sqrt{3}}{\varepsilon n}\right). \tag{3.42}$$

To complete the proof, we need to choose h_1 and u such that

$$\frac{h_1}{\varsigma\sqrt{n}} = o(1), \quad \frac{u}{\varsigma\sqrt{n}} = o(1), \quad \frac{\varsigma}{u} = o(1), \quad \text{and} \quad h_1^{-1}\varsigma\varepsilon n^{3/2} = o(1). \tag{3.43}$$

It is straightforward to verify that relations in (3.43) hold with $h_1 = \varsigma^{1+\frac{3\beta}{2}}$ and $u = \varsigma^{1-\frac{\beta_1}{2}}$, $\beta_{1,2} > 0$, $\beta_1 + \beta_2 < 2\alpha/3$, and n as in (3.34). □

4. The Poincaré map. In the present section, we consider the type I model, i.e., the randomly perturbed system with the stochastic forcing acting via the fast subsystem (see (2.9) and (2.10)). In the active phase of bursting (when the system undergoes spiking), the trajectory of the randomly perturbed system remains in the vicinity of the cylinder foliated by the periodic orbits of the fast subsystems (see Figure 6a). The time that the trajectory spends near L determines the duration of the active phase. The goal of this section is to describe the slow dynamics near L. In particular, we will estimate the distribution of the number of spikes in one burst. To this end, we introduce a transverse to L cross-section Σ (see Figure 6a) and construct the first return map. Specifically, we estimate the change in the state of the system after one cycle of rotation of the trajectory around L. The construction of the first return map for (2.9) and (2.10) is done in analogy to that for the deterministic models of bursting (see [34, 31]). However, the treatment of the randomly perturbed system requires certain modifications. First, we have to resolve the ambiguity in the notion of the first return time. The latter is due to the fact that generically a trajectory of the randomly perturbed system makes multiple crossings with Σ during each cycle around L. We refer the reader to the comments following Theorem 2.3 in [19] for an explicit example illustrating this effect. For the randomly perturbed system, we define the time of the first return so that it approaches the first return time of the underlying deterministic system in the limit of vanishing random perturbation. The definition of the first return time motivates the definition of the Poincaré map (see Definition 4.2). In sections 4.1 and 4.2, we use asymptotic expansions to construct the linear approximation for the Poincaré map of the fast subsystem. Here, we use an obvious observation that on finite time intervals and for sufficiently small $\epsilon > 0$, the slow variable typically remains in an $O(\epsilon)$ neighborhood of its initial value. Therefore, for finite times the Poincaré map of the fast subsystem captures the dynamics of the full system. Since we are interested in long-term behavior of the system, to complete the description of the first return map we also need to track the (small) changes in the slow variable after each cycle of oscillations. This is done in section 4.3, where we derive a 1D map for the slow variable. The combination of the 1D Poincaré map for the fast subsystem and that for the slow variable provides the first return map for the full problem (2.9) and (2.10). The linear approximation of the 2D map is used in section 4.4 to estimate the distribution of the number of spikes in one burst for the type I model. Effectively, the problem is reduced to the exit problem for a 1D linear randomly perturbed map. For the latter problem, we have already developed necessary analytical tools in section 3. Finally, in section 4.5, we comment on the straightforward modifications necessary to extend the analysis of this section to cover type II models.

4.1. Preliminary transformations. Recall that Σ stands for the transverse section located as shown schematically in Figure 6a. Let $y_0 < y_{bp}$ be outside an $O(\sigma)$ neighborhood of y_{bp}, and let $x_0 = \left(x_0^1, x_0^2\right)^T \in \Sigma$ be from an $O(\sigma)$ neighborhood of L. Consider an initial value problem (IVP) for (2.9) and (2.10) with initial data (x_0, y_0).

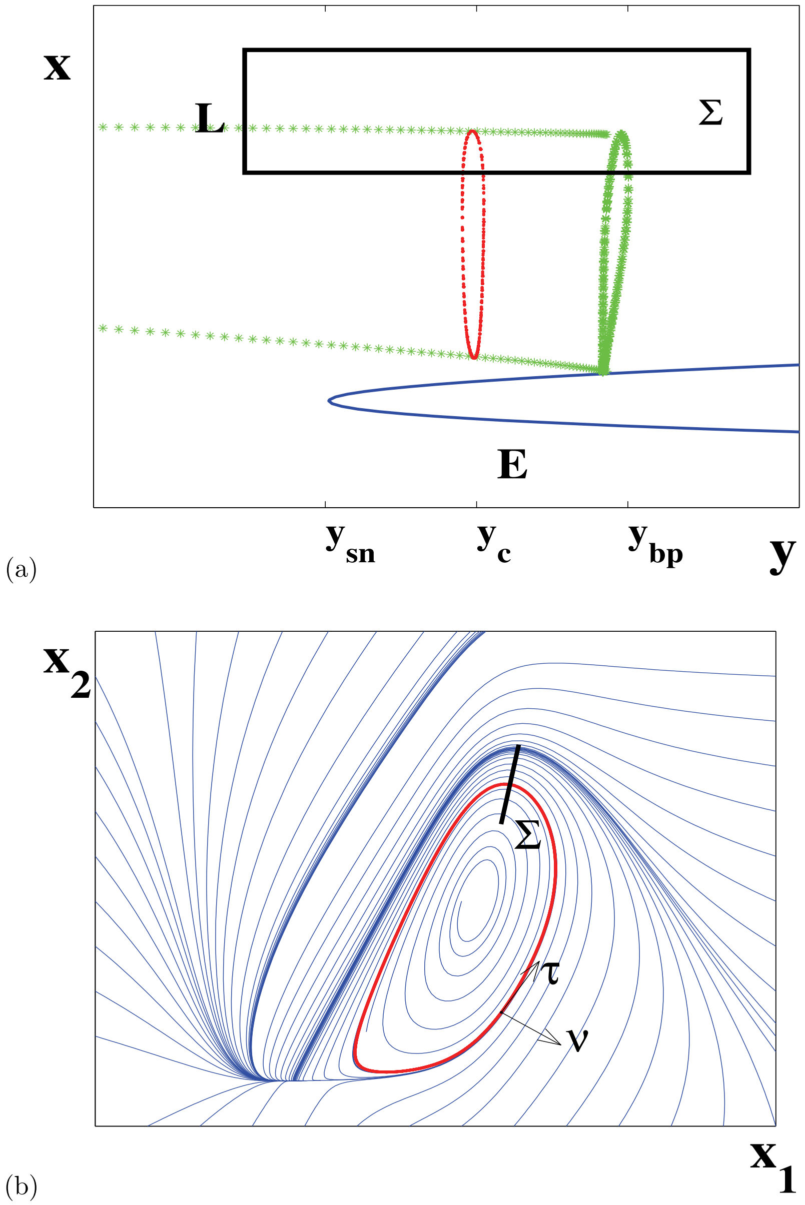


FIG. 6. (a) *Cross-section* Σ *is used in the construction of the first return map.* (b) *The phase plane of the fast subsystem* (2.1) *for* $y \in (y_{sn}, y_{bp})$.

By standard results from the asymptotic theory for randomly perturbed systems [19], we have the estimate

$$y_t = y_0 + O(\epsilon), \tag{4.1}$$

valid on a finite interval of time $t \in [0, \bar{t}]$. Here and below, for a small parameter $\epsilon > 0$, the symbols $O(\epsilon)$ and $o(\epsilon)$ in the asymptotic expansions of the random functions mean that the corresponding relations hold almost surely (a.s.). Specifically, $\psi_t(\epsilon) = O(\epsilon)$

for $t \in [t_1, t_2]$ means that there exists $\epsilon_0 > 0$ such that

$$\sup_{\substack{t \,\in\, [t_1,\, t_2] \\ \epsilon \,\in\, [0,\, \epsilon_0]}} \left|\epsilon^{-1}\psi_t(\epsilon)\right| < \infty \quad \text{a.s.}$$

In a similar fashion, we interpret $\psi_t(\epsilon) = o(\epsilon)$ when $\psi_t(\epsilon)$ is a random function.

By plugging (4.1) into (2.9), we obtain the following stochastic ODE:

$$dx_t = f\,(x_t)\,dt + \sigma p(x_t) dw_t + O(\epsilon), \tag{4.2}$$

where $f\,(x) := f\,(x, y_0)$, $p(x) := p\,(x, y_0)$, and y_0 is fixed. Equation (4.2) with $\epsilon = \sigma = 0$ has an exponentially orbitally stable periodic solution $x = \phi(t, y_0)$ of period $\mathcal{T}(y_0)$:

$$L(y_0) = \{x = \phi(\theta, y_0) :\ \theta \in [0, \mathcal{T}(y_0))\} \qquad \text{(cf. (2.2))}.$$

To simplify the notation, throughout the analysis of the fast subsystem, we will not explicitly indicate the dependence on y_0 when referring to L, ϕ, and $\mathcal{T}$. At each point $x = \phi(\theta) \in L$, we define vectors

$$\tau(\theta) = \left(f^1(x), f^2(x)\right)^T \quad \text{and} \quad \nu(\theta) = Jf(x), \ \text{ where } J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{4.3}$$

pointing in the tangential and normal directions, respectively. To study the trajectories of (4.2) in a small neighborhood of L, it is convenient to rewrite (4.2) in normal coordinates (θ, ξ) [23]:

$$x = \phi(\theta) + \xi\nu(\theta), \quad \theta \in [0, \mathcal{T}). \tag{4.4}$$

Lemma 4.1. *For sufficiently small $\delta > 0$, (4.4) defines a smooth change of coordinates in*

$$B_\delta = \{x = \phi(\theta) + \xi\nu(\theta) :\ |\xi| < \delta,\ \theta \in [0, \mathcal{T})\}. \tag{4.5}$$

In new coordinates, (4.2) has the form

$$d\theta_t = (1 + b_1(\theta_t)\xi_t)dt + \sigma h_1(\theta_t, \xi_t)\,(1 + b_2(\theta_t)\xi_t)\,dw_t + O(\epsilon, \delta^2, \sigma^2), \tag{4.6}$$

$$d\xi_t = a(\theta_t)\xi_t dt + \sigma h_2(\theta_t, \xi_t) dw_t + O(\epsilon, \delta^2, \sigma^2), \tag{4.7}$$

where smooth functions $a(\theta), b_1(\theta)$, and $b_2(\theta)$ are $\mathcal{T}$-periodic and

$$0 < \mu := \exp\left(\int_0^{\mathcal{T}} a(\theta) d\theta\right) = \exp\left(\int_0^{\mathcal{T}} \operatorname{div} f\,(\phi(\theta))\right) < 1, \tag{4.8}$$

$$h_1(\theta, \xi) = \frac{\langle p, \tau\rangle}{\langle \tau, \tau\rangle} = \frac{p^1 f^1 + p^2 f^2}{|f|^2},\ \ h_2(\theta, \xi) = \frac{\langle p, \nu\rangle}{\langle \tau, \tau\rangle} = \frac{p^2 f^1 - p^1 f^2}{|f|^2}. \tag{4.9}$$

Proof. The proof of the lemma follows along the lines of the proof of Theorem VI.1.2 in [23]. Let $z = (z^1, z^2)^T := (\theta, \xi)^T$ and denote the transformation in (4.4) by

$$x = v(z),\ z \in B_\delta. \tag{4.10}$$

Note

$$|Dv(\theta, 0)| = \begin{vmatrix} {\phi^1}'(\theta) & -f^2\,(\phi(\theta)) \\ {\phi^2}'(\theta) & f^1\,(\phi(\theta)) \end{vmatrix} = |f\,(\phi(\theta))|^2 \neq 0,\ \theta \in [0, \mathcal{T}).$$

Therefore, for sufficiently small $\delta > 0$, (4.10) defines a smooth invertible transformation in B_δ. Denote the inverse of v by $z = u(x)$, $x \in v(B_\delta)$, and note that

$$[Du(x)]^{-1} = Dv(z),\ x \in v(B_\delta). \tag{4.11}$$

By Itô's formula, we have

$$dz_t = Du(x_t)dx_t + O(\sigma^2)dt \tag{4.12}$$

and, therefore,

$$Dv(z_t)dz_t = dx_t + O(\sigma^2)dt. \tag{4.13}$$

By recalling that $z = (\theta, \xi)$ and after plugging (4.2) into (4.13), we obtain

$$\begin{aligned}\left[\frac{d\phi(\theta_t)}{d\theta} + \frac{d\nu(\theta_t)}{d\theta}\xi_t\right] d\theta_t + \nu(\theta_t)d\xi_t = \left(f\left(\phi(\theta_t)\right) + Df\left(\phi(\theta_t)\right)\nu(\theta_t)\xi_t + Q(\theta_t, \xi_t)\right) dt \\ + \sigma p dw_t + O(\epsilon, \sigma^2),\end{aligned} \tag{4.14}$$

where

$$Q(\theta, \xi) = f\left(\phi(\theta) + \xi\nu(\theta)\right) - f\left(\phi(\theta)\right) - Df\left(\phi(\theta)\right)\nu(\theta)\xi = O\left(\xi^2\right), \quad |\xi| < \delta.$$

Note that

$$\frac{d\phi(\theta)}{d\theta} = f\left(\phi(\theta)\right) = \tau(\theta), \quad \tau^T(\theta)\tau(\theta) = \nu(\theta)^T\nu(\theta) = \left|f\left(\phi(\theta)\right)\right|^2, \tag{4.15}$$

$$\frac{d\nu(\theta)}{d\theta} = \frac{d}{d\theta}Jf\left(\phi(\theta)\right) = JDf\left(\phi(\theta)\right)f\left(\phi(\theta)\right). \tag{4.16}$$

Taking into account (4.15) and (4.16), we project (4.14) onto the subspace spanned by $\tau(\theta_t)$ and after some algebra obtain

$$\dot{\theta}_t = 1 + \frac{f^TQ + f^T\left[DfJ - JDf\right]f\xi_t + \sigma f^Tp\dot{w}_t + O(\epsilon)}{f^Tf + f^TJDff\xi_t}. \tag{4.17}$$

Here and for the rest of the proof, for brevity we use the following notation:

$$f := f\left(\phi(\theta_t)\right), \quad Q := Q(\theta_t, \xi_t), \quad \text{and} \quad \nu := \nu(\theta_t).$$

Equation (4.17) can be rewritten as (4.6) with

$$\begin{aligned} b_1(\theta_t) &= \frac{1}{|f|^2}f^T\left[DfJ - JDf\right]f, \\ b_2(\theta_t) &= \frac{1}{|f|^2}f^TJDff. \end{aligned}$$

Similarly, by projecting (4.14) onto the subspace spanned by $\nu(\theta)$ and using (4.15) and (4.14), we derive

$$\dot{\xi}_t = a(\theta_t)\xi_t + \sigma h_2(\theta_t)\dot{w}_t + O\left(\delta^2\right),$$

where

$$a(\theta_t) = \frac{1}{|\nu|^2}\nu^T\left[Df\nu + \frac{d\nu}{d\theta}\right] - \frac{2\nu^T}{|\nu|^2}\frac{d\nu}{d\theta}.$$

The expression in the square brackets can be simplified as follows:

$$\nu^T \left[Df\nu + \frac{d\nu}{d\theta} \right] = f^T \left[J^T Df J + Df \right] = \mathrm{div} f\left(\phi(\theta)\right) |f|^2 .$$

Also,

$$\frac{2\nu^T}{|\nu|^2} \frac{d\nu}{d\theta} = \frac{2}{|f|^2} f^T J^T \frac{d}{d\theta} J f = \frac{2}{|f|^2} f^T \frac{d}{d\theta} f = \frac{1}{|f|^2} \frac{d}{d\theta} |f|^2 = \frac{d}{d\theta} \ln |f\left(\phi(\theta)\right)|^2 .$$

Therefore,

$$a(\theta) = \mathrm{div} f\left(\phi(\theta)\right) - \frac{d}{d\theta} \ln |f\left(\phi(\theta)\right)|^2 . \tag{4.18}$$

Equation (4.18) implies (4.8), since the integral over $[0, \mathcal{T}]$ of the last term on the right-hand side of (4.18) is zero. □

4.2. The Poincaré map for the fast subsystem. In the present subsection, we analyze the trajectories of the randomly perturbed system (4.2) lying close to the limit cycle $L(y_0)$, $y_0 < y_{bp}$. To this end, we consider an IVP for (4.6) and (4.7) subject to the initial condition

$$\theta_0 = 0 \quad \text{and} \quad |\xi_0| < \delta. \tag{4.19}$$

Throughout this section, we assume (even when it is not stated explicitly) that $\delta > 0$ is sufficiently small. It will be convenient to view the range of θ_t as $\mathbb{R}^1$ rather than a circle. Equation (4.4) provides the transformation of (θ_t, ξ_t) to the Cartesian coordinates even when θ_t exceeds $\mathcal{T}$.

We now turn to the construction of the Poincaré map. Condition $\theta = 0$ defines a transverse cross-section of $L(y_0)$, Σ. The trajectory of the deterministic system (4.6) and (4.7) with $\sigma = 0$ returns to Σ in time $\mathcal{T} + O(\xi_0)$. To define the Poincaré map for the randomly perturbed system, we also use another transverse cross-section $\tilde{\Sigma}$, which is located at an $O(1)$ distance away from Σ. Let (θ_t, ξ_t) be the solution of the IVP (4.6), (4.7), and (4.19) and

$$\tilde{T} = \inf\{t > 0 : \ (\theta_t, \xi_t) \in \tilde{\Sigma}\}.$$

Definition 4.2. *By the time of the first return of the trajectory* (4.6), (4.7), *and* (4.19) *to* Σ, *we call stopping time* T *such that*

$$T = \inf\{t > \tilde{T} : \ \theta_t = \mathcal{T}\}. \tag{4.20}$$

The first return map for (4.6), (4.7), *and* (4.19) *is defined as*

$$\bar{\xi} = P(\xi_0), \quad \textit{where} \quad \bar{\xi} = \xi_T.$$

In the remainder of this subsection, we compute the linear part of the Poincaré map. In the asymptotic expansions below, we do not indicate the dependence of the remainder terms on $\epsilon > 0$. The latter is assumed to be sufficiently small so that it has no effect on the leading order approximation of the Poincaré map.

The following notation is reserved for four functions, which will appear frequently in the asymptotic expansions below:

$$A(t,s) = \exp\left\{ \int_s^t a(u) du \right\}, \quad A(t) = A(t,0),$$
$$B(t,s) = \int_s^t A(u,s) b_1(u) du, \quad B(t) = B(t,0).$$

LEMMA 4.3. *On a finite time interval* $t \in [0, \bar{t}]$, $0 < \bar{t} < \infty$, *the solution of the IVP* (4.6), (4.7), *and* (4.19) *admits the asymptotic expansion*

$$\theta_t = \theta_t^{(0)} + \sigma\theta_t^{(1)} + O(\sigma^2, \xi_0^2), \tag{4.21}$$

$$\xi_t = \xi_t^{(0)} + \sigma\xi_t^{(1)} + O(\sigma^2, \xi_0^2). \tag{4.22}$$

The leading order coefficients are given by

$$\theta_t^{(0)} = t + \xi_0 B(t) + O(\xi_0^2), \tag{4.23}$$

$$\xi_t^{(0)} = \xi_0 A(t) + O(\xi_0^2). \tag{4.24}$$

The first order terms are given by the Gaussian diffusion process $z_t = (\theta_t^{(1)}, \xi_t^{(1)})^T$*:*

$$z_t = \int_0^t U(t,s)h(s)dw_s + O(\xi_0), \tag{4.25}$$

where

$$U(t,s) = \begin{pmatrix} 1 & B(t,s) \\ 0 & A(t,s) \end{pmatrix}, \quad h(t) := h(t,0) = (h_1(t,0), h_2(t,0))^T. \tag{4.26}$$

Proof. The procedure for constructing asymptotic expansions of solutions for a class of IVP, which includes (4.6), (4.7) and (4.19), can be found in [2, 19]. These sources also contain the estimates controlling the remainder terms. The coefficients $\theta_t^{(0,1)}$ and $\xi_t^{(0,1)}$ are determined as follows. By plugging (4.21) and (4.22) into (4.6) and (4.7) and extracting the coefficients multiplying different powers of σ, one obtains IVPs for the functions on the right-hand sides of (4.21) and (4.22). Specifically, for the leading order terms we have the following IVP:

$$\dot{\theta}_t^{(0)} = 1 + b_1\left(\theta_t^{(0)}\right)\xi_t^{(0)}, \tag{4.27}$$

$$\dot{\xi}_t^{(0)} = a\left(\theta_t^{(0)}\right)\xi_t^{(0)}, \tag{4.28}$$

$$\xi_0^{(0)} = \xi_0, \ \theta_0^{(0)} = 0. \tag{4.29}$$

To the next order,

$$\dot{z}_t = \Lambda(t, \xi_0)z_t + h\left(\theta_t^{(0)}, \xi_t^{(0)}\right)\dot{w}_s, \tag{4.30}$$

$$z_0 = 0, \tag{4.31}$$

where $z_t = \left(\theta_t^{(0)}, \xi_t^{(1)}\right)^T$, $h = (h_1, h_2)^T$, and

$$\Lambda(t, \xi_0) = \begin{pmatrix} b_1'\left(\theta_t^{(0)}(\xi_0)\right)\xi_t^{(0)}(\xi_0) & b_1\left(\theta_t^{(0)}(\xi_0)\right) \\ a'\left(\theta_t^{(0)}(\xi_0)\right)\xi_t^{(0)}(\xi_0) & a\left(\theta_t^{(0)}(\xi_0)\right) \end{pmatrix}. \tag{4.32}$$

Here, we explicitly indicated the dependence of the leading order coefficients on ξ_0 and used a prime to denote the differentiation with respect to θ. Formulae (4.23)–(4.26) in the statement of the lemma follow from (4.27)–(4.32). The details can be found in the appendix to this paper. ☐

Next, we calculate the time of the first return.

Lemma 4.4. *The time of the first return is given by*

$$T = T^{(0)} + \sigma T^{(1)} + o(\sigma) + O(\xi_0^2), \tag{4.33}$$

where

$$T^{(0)} = \mathcal{T} - \xi_0 B(\mathcal{T}) + O(\xi_0^2), \tag{4.34}$$

$$T^{(1)} = -\sigma\theta^{(1)}_{\mathcal{T}} = -\sigma \int_0^{\mathcal{T}} \left[h_1(u) + B(\mathcal{T}, u) h_2(u)\right] dw_u. \tag{4.35}$$

Proof. From the definition of the first return time, (4.21), and (4.23), we have

$$T + \xi_0 B(T) + \sigma\theta^{(1)}_T + O(\sigma^2, \xi_0^2) = \mathcal{T} \text{ a.s.} \tag{4.36}$$

Thus,

$$\lim_{\sigma\to 0} T = T^{(0)}(\xi_0) \text{ a.s.}, \tag{4.37}$$

where $T^{(0)}(\xi_0)$ is found from the following equation:

$$T^{(0)}(\xi_0) + \xi_0 B\left(T^{(0)}(\xi_0)\right) + O(\xi_0^2) = \mathcal{T}. \tag{4.38}$$

Equation (4.38) implies (4.34). Furthermore, the combination of (4.34), (4.36), and (4.37) yields (4.35). □

Lemma 4.5. *The first return map is given by the*

$$\bar{\xi} = \mu\xi\left(1 + \sigma r_1\right) + \sigma r_2 + o(\sigma) + O(\xi_0^2), \tag{4.39}$$

where Gaussian RVs $r_{1,2}$ are given by

$$r_1 = -a(0) \int_0^{\mathcal{T}} \left[h_1(u) + B(\mathcal{T}, u) h_2(u)\right] dw_u, \quad r_2 = \int_0^{\mathcal{T}} A(\mathcal{T}, u) h_2(u) dw_u. \tag{4.40}$$

Proof. From (4.22), (4.24)–(4.26), and (4.33), we have

$$\begin{aligned} \bar{\xi} = \xi_T &= \xi_0 A(T) + \sigma \int_0^T A(T, s) h_2(s) dw_s + O(\sigma^2, \xi_0^2) \\ &= \xi_0 A(T) + \sigma r_2 + O(\sigma^2, \xi_0^2), \end{aligned} \tag{4.41}$$

where r_2 is defined in (4.40). The first term on the right-hand side of (4.41) can be rewritten as follows:

$$\begin{aligned} A(T) = A(\mathcal{T}) A(\mathcal{T} + \sigma T^{(1)}, \mathcal{T}) + o(\sigma) + O(\xi_0) &= \mu \exp\left(\sigma a(0) T^{(1)}\right) + o(\sigma) \\ = \mu\left(1 - \sigma a(0)\theta^{(1)}_{\mathcal{T}}\right) + o(\sigma) + O(\xi_0). \end{aligned} \tag{4.42}$$

Finally, we extract the expression for $\theta^{(1)}_{\mathcal{T}}$ from (4.25) and (4.26):

$$\theta^{(1)}_{\mathcal{T}} = \int_0^{\mathcal{T}} \left[h_1(u) + B(\mathcal{T}, u) h_2(u)\right] dw_u. \tag{4.43}$$

Equations (4.41)–(4.43) yield (4.39) and (4.40). □

Remark 4.6. We close this section by observing that, as follows from (4.40), RVs r_1 and r_2 are stochastic integrals of different deterministic functions, say $f(t)$ and $g(t)$, with respect to the same Brownian motion over the interval $[0, \mathcal{T}]$. Consequently, their joint distribution is bivariate normal with 0 mean vector and a covariance matrix whose diagonal entries are

$$\int_0^{\mathcal{T}} f^2(t)dt \quad \text{and} \quad \int_0^{\mathcal{T}} g^2(t)dt,$$

and whose off-diagonal entry is

$$\int_0^{\mathcal{T}} f(t)g(t)dt.$$

This is perhaps easiest to see by using Riemann representation of a stochastic integral (see, e.g., [42, Proposition 7.6]), basic properties of Brownian motion, and the fact that a random vector is multivariate normal iff any linear combination of its components is a normal RV.

4.3. The first return map for the slow variable.

Our next goal is to estimate the change of the slow variable, y_t, after one cycle of oscillations of the fast subsystem for the following initial conditions:

$$0 < y_{bp} - y_0 = O(1),\ x_0 = \phi(0) + \xi_0\nu(0) \in \Sigma, \ \text{ and } |\xi_0| < \delta. \tag{4.44}$$

We denote the first return map for y by

$$\bar{y} = P(y, \xi_0), \quad \text{where} \quad P(y_0, \xi_0) = y_T,$$

and T is the first return time of the fast subsystem (see Definition 4.2).

LEMMA 4.7. *The first return map for* y *has the following form:*

$$P(y, \xi) = y + \epsilon G(y) + \epsilon\sigma r_3 + \epsilon a\xi + o(\epsilon\sigma), \tag{4.45}$$

where

$$G(y) = \int_0^{\mathcal{T}} g\left(\phi(s), y\right) ds \tag{4.46}$$

and $r_3 = N\left(0, O(1)\right)$ *and* a *is a constant independent of* σ *and* ϵ.

Remark 4.8. Recall that $\mathcal{T}$ and $\phi(\cdot)$ are functions of slow variable y (see (2.2)). To avoid using cumbersome notation we continue to suppress the dependence on y.

Proof of Lemma 4.7. By (2.10),

$$y_T = y_0 + \epsilon \int_0^T g(x_s, y_0)ds + O(\epsilon^2), \tag{4.47}$$

where x_s satisfies IVP (4.6), (4.7), and (4.19). Let $x = \phi(\theta) + \xi\nu(\theta)$ and denote

$$\tilde{g}(\theta, \xi, y) := g(x, y),\ g_0(s) = \tilde{g}(s, 0),\ g_1(s) = \frac{\partial \tilde{g}}{\partial \theta}(s, 0), \text{ and } g_2(s) = \frac{\partial \tilde{g}}{\partial \xi}(s, 0). \tag{4.48}$$

Using (4.48), we rewrite (4.47) as

$$y_T = y_0 + \epsilon \int_0^T \tilde{g}(\theta_s^{(0)} + \sigma\theta_s^{(1)}, \xi_s^{(0)} + \sigma\xi^{(1)}) + O(\epsilon\sigma^2). \tag{4.49}$$

Using the Taylor expansion for $\tilde{g}$ in (4.49) and (4.21), (4.22) and (4.33), from (4.49) we derive

$$y_T = y_0 + \epsilon \int_0^{\mathcal{T}} \left\{ g_0(s) + g_1(s) \left[\xi_0 B(s) + \sigma \theta_s^{(1)} \right] + g_2(s) \left[\xi_0 A(s) + \sigma \xi_s^{(1)} \right] \right\} ds$$

$$+ \int_{\mathcal{T}}^{\mathcal{T} - \xi_0 B(\mathcal{T}) - \sigma \theta_{\mathcal{T}}^{(1)}} g_0(s) ds + o(\epsilon\sigma) + O(\epsilon \xi_0^2). \tag{4.50}$$

We approximate the last integral on the right-hand side of (4.50) by

$$\int_{\mathcal{T}}^{\mathcal{T} - \xi_0 B(\mathcal{T}) - \sigma \theta_{\mathcal{T}}^{(1)}} g_0(s) ds = -g_0(0) \left[\xi_0 B(\mathcal{T}) + \sigma \theta^{(1)} \right] + o(\sigma, \xi_0). \tag{4.51}$$

The combination of (4.50) and (4.51) implies (4.45) with

$$a = \int_0^{\mathcal{T}} \left[g_1(s) B(s) + g_2(s) A(s) \right] ds - g_0(0) B(\mathcal{T}), \tag{4.52}$$

$$r_3 = \int_0^{\mathcal{T}} \left[g_1(s) \theta_s^{(1)} + g_2(s) \xi_s^{(1)} \right] ds. \qquad \square \tag{4.53}$$

4.4. The exit problem. In this subsection, we first combine the return maps derived for the slow and fast subsystems to obtain the Poincaré map for the full 3D system. Next, we approximate the Poincaré map and the BA of the limit cycle $L(y_c)$ and characterize the distribution of the exit times for the approximate problem. This distribution is then related to the distribution of the number of spikes within bursting episodes. To approximate the Poincaré map we linearize it around the stable fixed point of the deterministic map corresponding to the limit cycle $L(y_c)$. Aside from the systematic derivation of the Poincaré map in the previous subsections, we offer no rigorous justification for substituting the nonlinear Poincaré map with its linear part in the analysis of the exit problem. While in general such approximation may not be accurate, we believe that for the present problem the analysis of the linearized system captures the statistics of the first exit times well for the following reason. In models of square-wave bursting the limit cycle generating spiking is often located close to the boundary of its BA (see Figure 6b for a representative example). Therefore, before the trajectories leave the BA, they remain in a small neighborhood of the limit cycle, where the linear part of the vector field governs the dynamics. After these preliminary remarks, we turn to the derivation of the approximate problem and its analysis.

Lemmas 4.5 and 4.7 yield the asymptotic formulae for the first return map of the randomly perturbed system (2.9) and (2.10) in the normal coordinates (4.4):

$$\xi_{n+1} = \mu \xi_n \left(1 + \sigma r_{1,n}\right) + \sigma r_{2,n} + o(\sigma), \tag{4.54}$$

$$y_{n+1} = y_n + \epsilon G(y_n) + \epsilon \sigma r_{3,n} + \epsilon a \xi_n + o(\epsilon\sigma),\ n = 0, 1, 2, \ldots, \tag{4.55}$$

where (ξ_0, y_0) are given in (4.44) and the expressions for a and $r_{i,n},\ i = 1, 2, 3$, are are given in (4.40), (4.52), and (4.53). Recall that by (SS) (see section 2), $G(y)$ has a simple zero at $y = y_c$ and $\lambda := -G'(y_c) > 0$. Thus, $(0, y_c)$ is an attracting fixed point of the unperturbed map (4.54) and (4.55) with $\sigma = 0$. The linearization of (4.54) and (4.55) about $(0, y_c)$ yields

$$\xi_{n+1} = \mu \xi_n \left(1 + \sigma \tilde{r}_{1,n}\right) + \sigma \tilde{r}_{2,n}, \tag{4.56}$$

$$\eta_{n+1} = \lambda \eta_n + \epsilon \sigma \tilde{r}_{3,n} + \epsilon a_2 \xi_n, \quad n = 0, 1, 2, \ldots, \tag{4.57}$$

where $\eta = y - y_c$, $0 < \lambda = 1 - \epsilon a_1$, and $0 < \mu < 1$. The distributions of the RVs $r_{i,n},\ i = 1,2,3$, depend on y_n, as both the upper bound of integration $\mathcal{T}$ and the integrands in (4.40) and (4.53) are smooth functions of y. The stochastic terms $\tilde{r}_{i,n},\ i = 1,2,3$, in the linearized system are obtained by evaluating the expressions for $\tilde{r}_{i,n},\ i = 1,2,3$, in (4.40) and (4.53) at $y = y_c$. Thus, $(\tilde{r}_{1,n}, \tilde{r}_{2,n}, \tilde{r}_{3,n})$ are IID copies of an $N(0, \Sigma_3)$, where the entries of Σ_3 are $O(1)$ in a sense that they do not depend on any other parameters. Further, we approximate the BA of $L(y_c)$ by a cylindrical shell, so that in the (ξ, η) coordinate plane, it projects to $\Pi := \left[-\tilde{h}_\xi, h_\xi\right] \times \left[-\tilde{h}_\eta, h_\eta\right]$ for some $\tilde{h}_{\xi,\eta} > h_{\xi,\eta} > 0$ independent of $\sigma > 0$. Each iteration of the Poincaré map corresponds to a spike within a burst. The burst terminates when the trajectory leaves the BA of $L(y_c)$. Assuming that the linearization (4.56) and (4.57) and Π provide suitable approximations for the Poincaré map and the BA of $L(y_c)$, respectively, the distribution of the number of spikes in one burst can be approximated by the distribution of the first exit times for the trajectories of (4.56) and (4.57) from Π:

$$\tau = \min\{\tau_\xi, \tau_\eta\}, \tag{4.58}$$

where

$$\tau_\xi = \inf_{n>0}\{\xi_n > h_\xi\} \quad \text{and} \quad \tau_\eta = \inf_{n>0}\{\eta_n > h_\eta\}.$$

We are now in a position to apply the results of section 3 to describe the distribution of (4.58). By Theorem 3.7, the distribution of τ is asymptotically geometric with parameter

$$p \approx \frac{\sigma}{C\sqrt{2\pi}} e^{-\frac{C}{\sigma^2}} \tag{4.59}$$

for some $C > 0$ independent of ϵ and σ. In the proof of Theorem 3.7, we studied a class of 2D randomly perturbed maps that includes (4.56) and (4.57). However, the distribution of τ is effectively determined by the first equation (4.56), i.e., by the 1D first return map of the fast subsystem. This can be seen by observing that according to the approximations given at the end of the proof of Theorem 3.7 (see the arguments following (3.30)) if $\epsilon > 0$ is sufficiently small, then $\tau_\xi \ll \tau_\eta$ and $\tau \sim \tau_\xi$. Thus, in type I models the distribution of spikes in one burst is effectively determined by the 1D first return map for the fast subsystem (4.56). In particular, the statistics of the number of spikes in one burst does not depend on the relaxation parameter $\epsilon > 0$, provided the latter is sufficiently small.

4.5. Type II model. The derivation of the Poincaré map for the type II models differs from the analysis in sections 4.1–4.4 for type I models only in some minor details. In this subsection, we comment on the necessary modifications and state the final result. Recall that in contrast to type I models, in (2.11) and (2.12), stochastic forcing enters the slow equation. As before, the initial condition is given by (4.44). On finite time intervals, solutions of the IVP for (2.11) and (2.12) admit the following asymptotic expansions:

$$x_t = x_t^{(0)} + \epsilon\sigma x_t^{(1)} + O\left((\epsilon\sigma)^2\right), \tag{4.60}$$

$$y_t = y_t^{(0)} + \epsilon\sigma y_t^{(1)} + O\left((\epsilon\sigma)^2\right), \tag{4.61}$$

where the first order corrections $x_t^{(1)}$ and $y_t^{(1)}$ are Gaussian processes (cf. Theorem 2.2 in [19]). Using (4.60) and (4.61), we obtain the leading order approximation of the fast subsystem:

$$\dot{x}_t = f(x_t, y_0) + \epsilon\sigma \frac{\partial f(x_t^{(0)}, y_0)}{\partial y} y_t^{(1)} + o(\epsilon\sigma). \tag{4.62}$$

From this point, the derivation of the Poincaré map follows along the same lines as described in detail for type I models in sections 4.1–4.4. We omit any further details and state the final result, the linear approximation of the Poincaré map for the present case:

$$\xi_{n+1} = \mu\xi_n \left(1 + \epsilon\sigma\tilde{r}_{1,n}\right) + \epsilon\sigma\tilde{r}_{2,n}, \tag{4.63}$$

$$\eta_{n+1} = \lambda\eta_n + \epsilon\sigma\tilde{r}_{3,n} + \epsilon a_2\xi_n, \quad n = 0, 1, 2, \dots. \tag{4.64}$$

As in the previous case, we are interested in the distribution of the first exit time τ (see (4.58)). To estimate the latter, we use the same argument as in the previous subsection. This time the system is described by

$$\Theta_{n+1} = A_{n+1}\Theta_n + \sigma\epsilon G_{n+1}, \quad n \geq 1, \tag{4.65}$$

where A_n is as before and $G_n = \left[\begin{smallmatrix} r_{2,n} \\ r_{3,n} \end{smallmatrix}\right]$. The presence of the factor ϵ in both components of G_n leads to the following expression for the numerator of p (see (3.27)):

$$\mathbb{P}\left(\frac{\mu X_1 r_1 + r_2}{\sigma_{X_1}} > \frac{h_1 - \mu X_1}{\epsilon\sigma\sigma_{X_1}}, r_3 > \frac{h_2 - a_2 X_1}{\sigma} + \frac{\lambda(h_2 - X_2)}{\epsilon\sigma}, (X_1, X_2) \in B_h \right).$$

This expression decays very fast as a function of $h_2 - X_2$, and since X_2 has heavy tails it is approximated (up to inessential polynomial factors) by

$$\mathbb{P}\left(\frac{\mu X_1 r_1 + r_2}{\sigma_{X_1}} > \frac{h_1 - \mu X_1}{\epsilon\sigma\sigma_{X_1}}, r_3 > \frac{h_2 - a_2 X_1}{\sigma}, (X_1, X_2) \in B_h \right).$$

We are now in the analogous situation to that encountered in (3.28), except that the small parameter $\epsilon > 0$ appears in the denominator of the other variable. As a consequence, this time we obtain that $\tau_\xi \ll \tau_\eta$ for small $\epsilon > 0$. Therefore, in contrast to type I models, the escape of a trajectory of (2.11) and (2.12) from $\mathcal{A}$ is dominated by the slow subsystem, i.e., $\tau = \tau_\eta$.

5. Numerical example. In the present section, we illustrate the statistical regimes identified in this study with numerical simulations of a conductance-based model of a neuron in the presence of noise. To this end, we use a three-variable model of a bursting neuron introduced by Izhikevich [26]. The model dynamics is driven by the interplay of the three ionic currents: persistent sodium, I_{NaP}, the delayed rectifier, I_K, a slow potassium M-current, I_{KM}, and a passive leak current, I_L. The following system of three differential equations describes the dynamics of the membrane potential, v, and two gating variables, n and y:

$$C\dot{v} = F(v, n, y), \tag{5.1}$$

$$\tau_n \dot{n} = n_\infty(v) - n, \tag{5.2}$$

$$\tau_y \dot{y} = y_\infty(v) - y, \tag{5.3}$$

where $F(v, n, y) = -g_{NaP} m_\infty(v)(v - E_{NaP}) - g_K n(v - E_K) - g_{KM} y(v - E_K) - g_L(v - E_L) + I$; g_s and E_s, are the maximal conductance and the reversal potential

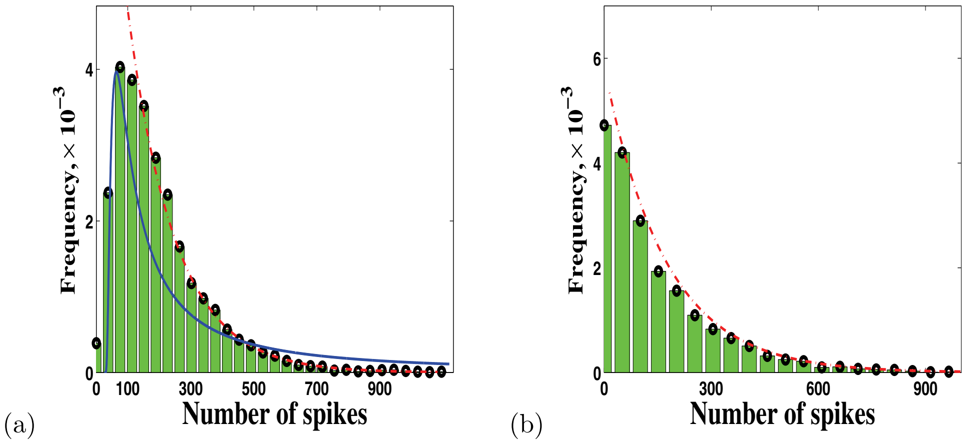


FIG. 7. *The histograms for the number of spikes in one burst. The histograms computed for the type I model in* (a) *and type II in* (b) *are normalized to approximate the corresponding PDFs. The tails of both functions are well approximated by the exponential densities with parameters* 0.0067 *and* 0.0125*, respectively. In* (b) *the exponential distribution already gives a very good approximation for the number of spikes exceeding* 10*. The region of exponential behavior in* (a) *starts around* $n \sim 100$*. In* (a)*, we also plotted in solid line the shifted diffusive density* $\Psi_a(x-25), a \approx 10.8$*. Although it is hard to claim a quantitative fit of the diffusive density and the data, the qualitative similarity between the diffusive pdf* $\Psi_a(x)$ *and the peak in the data in the range* $n \sim 50-100$ *is apparent. The values of parameters are* $C = 1$ $\left(\mu F cm^{-2}\right)$; $g_{NA} = 20$, $g_K = 10$, $g_{KM} = 5$, $g_L = 8$ $\left(mS cm^{-2}\right)$; $E_{Na} = 60$, $E_K = -90$, $E_L = -80$ (mV); $a_m = -20$, $a_n = -25$, $a_y = -10$ (mV); $b_m = 15$, $b_n = 5$, $b_y = 5$; $\tau_n = 0.152$, $\tau_y = 20$ $\left(ms^{-1}\right)$, $I = 5pA$*, and* $\sigma = 1$.

of I_s, $s \in \{NaP, K, KM, L\}$, respectively; and I is the applied current. The time constants τ_n and τ_y determine the rates of activation in the populations of K and KM channels. The steady-state functions are defined by $s_\infty(v) = \left(1 + \exp(\frac{a_s - v}{b_s})\right)^{-1}$, $s \in \{m, n, y\}$. The parameter values are given in the caption of Figure 7. This completes the description of the deterministic model. The random perturbation is used in the form of white noise, $\sigma \dot{w}_t$, and is added to the first equation (5.1) for the type I model or to the third one (5.3) for the type II model. After suitable rescaling, these models can be put in the nondimensional form (2.9), (2.10) or (2.11), (2.12). The separation of the timescales in the nondimesional models (i.e., small $\epsilon > 0$) is the result of the presence of the disparate time constants $\tau_h \gg \tau_n$ in the original model (see caption of Figure 7).

The parameters of the deterministic system are chosen so that it has a limit cycle located as shown in Figure 3c. In the presence of small noise the system generates bursting. In each numerical experiment, we integrated the randomly perturbed system using the Euler–Maruyama method [24] until it generated 5,000 bursts. We used these data to estimate the probability density for the number of spikes within one burst. In Figure 7, we plot the histograms for the number of spikes in one burst for type I and type II models. The histograms in Figure 7 are scaled to approximate the probability density function (PDF) for the number of spikes in one burst. Both PDFs shown in Figure 7a,b have distinct exponential tails as expected for the asymptotically geometric RVs. Note that the distribution in Figure 7a fits well with the geometric distribution for $N > 100$, while in Figure 7b the geometric distribution fits the data almost on the entire domain $N > 10$. In addition, the peak in the histogram in Figure 7a is reminiscent of the PDF characteristic for the diffusive escape (see Figure 5). For comparison, we plotted a slightly shifted diffusive PDF, $\Psi_a(x),\ a = 10.8$, in Figure 7a. Matching the data and Ψ_a is a delicate matter, because it is not clear how wide the range of n is, to which the estimates of Theorem 3.9 apply. Nonetheless, the qualitative similarity of the peak in the histogram in the range $n \sim 50 - 100$ and the diffusive PDF is apparent. We repeated these numerical experiments for a few other sets of parameters and found qualitatively similar results.

Collecting the statistical data shown in Figure 7 requires integrating the system over very long intervals of time, for which it would be hard to justify the accuracy of the Euler–Maruyama method. However, capturing the statistical features of the dynamical patterns does not require having an accurate solution on the entire interval of time, because they are determined by the discrete dynamics of the first return map. The iterations of the latter are expected to be insensitive to the numerical noise, as suggested by the analysis of the randomly perturbed maps in section 3. Therefore, we only need to have accurate numerical solutions on the time intervals comparable with the typical periods of the fast oscillations. This is easy to achieve with the Euler–Maruyama method. We repeated these numerical experiments using the second order Runge–Kutta method and obtained very similar results. These informal arguments form the rationale for using the above numerical scheme. The rigorous justification of the numerics is beyond the scope of this paper.

Appendix. In this appendix, we provide the details of the derivation of (4.23)–(4.26), which were omitted in the main part of the paper.

To derive (4.23) and (4.24), we first note that $\theta_t^{(0)}$ is a monotonic function on $[0,\bar{t}]$, provided $\delta > 0$ is sufficiently small. Thus,

$$\frac{d\xi^{(0)}}{d\theta^{(0)}} = a(\theta^{(0)})\xi^{(0)} + O(\xi_0^2)$$

and

$$\xi^{(0)}(\theta^{(0)}) = \xi_0 A(\theta^{(0)}) + O(\xi_0^2). \tag{A.1}$$

By plugging (A.1) into (4.27), we have

$$\dot{\theta}_t^{(0)} = 1 + b_1(\theta_t^{(0)})\xi_0 A(\theta^{(0)}). \tag{A.2}$$

By Gronwall's inequality,

$$\theta_t^{(0)} = \psi_t + O(\xi_0^2),\ t \in [0,\bar{t}], \tag{A.3}$$

where ψ_t solves

$$\dot{\psi}_t^{(0)} = 1 + \xi_0 b_1(t) A(t),\ \psi_0 = 0. \tag{A.4}$$

The combination of (A.1), (A.3), and (A.4) implies (4.24).

We next turn to IVP (4.30), (4.31), and (4.24). Let $U(t,\xi_0)$ denote the principal matrix solution of the homogeneous system

$$\dot{z}_t = \Lambda(t,\xi_0) z_t. \tag{A.5}$$

Then the solution of (4.30) and (4.31) is given by

$$z_t = \int_0^t U(t,s,\xi_0) h\left(\theta_s^{(0)}, \xi_s^{(0)}\right) dw_s = \int_0^t U(t,s) h\,(s,0)\, dw_s + O(\xi_0),\ t \in [0,\bar{t}], \tag{A.6}$$

where

$$U(t,s,\xi_0)=U(t,\xi_0)U^{-1}(s,\xi_0) \quad \text{and} \quad U(t,s)=U(t,s,0). \tag{A.7}$$

Finally, by integrating (A.5) with $\xi_0 = 0$ and appropriate initial conditions, one computes

$$U(t,0)=\begin{pmatrix} 1 & B(t) \\ 0 & A(t) \end{pmatrix}. \tag{A.8}$$

The expression for $U(t,s)$ in (4.26) follows from (A.7) and (A.8).

REFERENCES

[1] J. P. Baltanas and J. M. Casado, *Bursting behavior, of the FitzHugh–Nagumo neuron model subject to monochromatic noise*, Phys. D, 122 (1998), pp. 231–240.

[2] Yu. N. Blagoveshchenskii, *Diffusion processes depending on small parameter*, Theory Probab. Appl., 7 (1962), pp. 130–146.

[3] N. Berglund and B. Gentz, *Noise-Induced Phenomena in Slow-Fast Dynamical Systems: A Sample-Paths Approach*, Springer, London, 2006.

[4] J. Best, A. Borisyuk, J. Rubin, D. Terman, and M. Wechselberger, *The dynamic range of bursting in a model respiratory pacemaker network*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 1107–1139.

[5] R. J. Butera, J. Rinzel, and J. C. Smith, *Models of respiratory rhythm generation in the pre-Botzinger complex:* I. *Bursting pacemaker neurons*, J. Neurophysiol., 82 (1999), pp. 382–397.

[6] T. R. Chay, *Chaos in a three-variable model of an excitable cell*, Phys. D, 16 (1985), pp. 233–242.

[7] T. R. Chay and J. Rinzel, *Bursting, beating, and chaos in an excitable membrane model*, Biophys. J., 47 (1985), pp. 357–366.

[8] C. Chow and J. White, *Spontaneous action potentials due to channel fluctuations*, Biophys. J., 71 (1996), pp. 3013–3021.

[9] J. J. Collins, C. C. Chow, and T. T. Imhoff, *Aperiodic stochastic resonance in excitable systems*, Phys. Rev. E, 52 (1995), pp. R3321–R3324.

[10] R. E. L. DeVille, C. Muratov, and E. Vanden-Eijnden, *Two distinct mechanisms of coherence in randomly perturbed dynamical systems*, Phys. Rev. E, 72 (2005), article 031105.

[11] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1990. Reprint of the 1953 original.

[12] P. Embrechts and C. M. Goldie, *Perpetuities and random equations*, in Asymptotic Statistics, Physica, Heidelberg, 1994, pp. 75–86.

[13] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I., 3rd ed., Wiley, New York, 1968.

[14] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971/1972), pp. 193–226.

[15] R. F. Fox, *Stochastic versions of the Hodgkin–Huxley equations*, Biophys. J., 72 (1997), pp. 2068–2074.

[16] R. F. Fox and Y. Lu, *Emergent collective behavior in large numbers of globally coupled independently stochastic ion channels*, Phys. Rev. E, 49 (1994), pp. 3421–3431.

[17] M. I. Freidlin, *On stable oscillations and equilibriums induced by small noise*, J. Statist. Phys., 103 (2001), pp. 283–300.

[18] M. I. Freidlin, *On stochastic perturbations of dynamical systems with fast and slow components*, Stoch. Dyn., 1 (2001), pp. 261–281.

[19] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd ed., Springer, New York, 1998.

[20] R. M. Ghigliazza and P. Holmes, *Minimal models of bursting neurons: How multiple currents, conductances, and timescales affect bifurcation diagrams*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 636–670.

[21] C. M. Goldie, *Implicit renewal theory and tails of solutions of random equations*, Ann. Appl. Probab., 1 (1991), pp. 126–166.

[22] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, New York, 1983.

[23] J. Hale, *Ordinary Differential Equations*, 2nd ed., R. E. Krieger, Huntington, NY, 1980.

[24] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Rev., 43 (2001), pp. 525–546.

[25] A. A. Hill, J. Lu, M. A. Massino, O. H. Olsen, and R. L. Calabrese, *A model of a segmental oscillator in the leech heartbeat neuronal network*, J. Comput. Neurosci., 10 (2001), pp. 281–302.

[26] E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Boston, MA, 2007.

[27] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lecture Notes in Math. 1609, Springer, Berlin, 1995, pp. 44–118.

[28] N. L. Johnson and S. Kotz, *Discrete Distributions*, Wiley, New York, 1969.

[29] H. Kesten, *Random difference equations and renewal theory for products of random matrices*, Acta Math., 131 (1973), pp. 207–248.

[30] S. Kwapień and W. A. Woyczyński, *Random Series and Stochastic Integrals: Single and Multiple*, Birkhäuser, Basel, 1992.

[31] E. Lee and D. Terman, *Uniqueness and stability of periodic bursting solutions*, J. Differential Equations, 158 (1999), pp. 48–78.

[32] A. Longtin and K. Hinzer, *Encoding with bursting, subthreshold oscillations, and noise in mammalian cold receptors*, Neural Comput., 8 (1996), pp. 215–255.

[33] G. S. Medvedev, *Transition to bursting via deterministic chaos*, Phys. Rev. Lett., 97 (2006), article 048102.

[34] G. S. Medvedev, *Reduction of a model of an excitable cell to a one-dimensional map*, Phys. D, 202 (2005), pp. 37–59.

[35] G. S. Medvedev and J. E. Cisternas, *Multimodal regimes in a compartmental model of the dopamine neuron*, Phys. D, 194 (2004), pp. 333–356.

[36] L. S. Pontryagin and L. V. Rodygin, *Approximate solution of a system of ordinary differential equations involving a small parameter in the derivatives*, Soviet. Math. Dokl., 1 (1960), pp. 237–240.

[37] J. Rinzel, *A formal classification of bursting mechanisms in excitable systems*, in Proceedings of the International Congress of Mathematicians, A. M. Gleason, ed., AMS, Providence, RI, 1987, pp. 135–169.

[38] J. Rinzel and G. B. Ermentrout, *Analysis of neural excitability and oscillations*, in Methods in Neuronal Modeling, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1989.

[39] P. F. Rowat and R. C. Elson, *State-dependent effects of Na channel noise on neuronal burst generation*, J. Comput. Neurosci., 16 (2004), pp. 87–112.

[40] J. Rinzel and W. C. Troy, *A one-variable map analysis of bursting in the Belousov–Zhabotinskii reaction*, in Nonlinear Partial Differential Equations, J. A. Smoller, ed., AMS, Providence, RI, 1982, pp. 411–427.

[41] G. Smith, *Modeling the stochastic gating of ion channels*, in Computational Cell Biology, Interdiscip. Appl. Math. 20, C. P. Fall et al., eds., Springer, New York, 2002.

[42] J.M. Steele, *Stochastic Calculus and Financial Applications*, Springer, New York, 2001.

[43] J. Su, J. Rubin, and D. Terman, *Effects of noise on elliptic bursters*, Nonlinearity, 17 (2004), pp. 133–157.

[44] D. Terman, *The transition from bursting to continuous spiking in excitable membrane models*, J. Nonlinear Sci., 2 (1992), pp. 135–182.

[45] W. Vervaat, *On a stochastic difference equation and a representation of nonnegative infinitely divisible random variables*, Adv. in Appl. Probab., 11 (1979), pp. 750–783.

[46] J. White, J. Rubenstein, and A. Kay, *Channel noise in neurons*, Trends in Neurosci., 23 (2000), pp. 131–137.

[47] V.A. Zorich, *Mathematical Analysis.* II, Springer, Berlin, 2004.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1393–1414

DERIVATION OF A CONTINUUM MODEL FOR THE LONG-RANGE ELASTIC INTERACTION ON STEPPED EPITAXIAL SURFACES IN $2+1$ DIMENSIONS[*]

HAOYUN XU[†] AND YANG XIANG[†]

Abstract. In heteroepitaxy, the mismatch of lattice constants in the crystal film and the substrate causes a misfit stress in the bulk of the film, driving the self-organization of the film surface into various nanostructures. Below roughening transition temperature, the epitaxial surface consists of terraces separated by atomic-height steps, and the misfit results in a long-range elastic interaction between surface steps. In this case, the surface morphology is determined by the motion of the steps, and the widely used continuum models for surface evolution above the roughening transition temperature do not apply directly. In this paper, we present a continuum model for this long-range elastic interaction on a stepped heteroepitaxial surface in $2+1$ dimensions. The continuum model is derived rigorously by taking the continuum limit from the discrete model for the interaction between steps, thus incorporating the discrete features of the stepped surface.

Key words. epitaxial growth, misfit, steps, continuum model, elasticity

AMS subject classifications. 35Q72, 74A50, 74A10

DOI. 10.1137/080712581

1. Introduction.

The study of stress-driven morphological evolution of surfaces in epitaxial growth has attracted extensive interest recently. These stress-driven self-assembled nanostructures exhibit novel electronic and optical properties, which have various potential applications in semiconductor industry [33, 9, 10, 29]. Many continuum models can be found in the literature on surface morphological evolution under elastic effects [2, 11, 43, 42, 8, 52, 4, 53, 41, 10, 54, 50]. In these models, the surfaces are modeled as continuously changed profiles without discrete structures, and thus they work only for surfaces above the roughening transition temperature.

Below the roughening transition temperature, a crystal surface forms facets and steps, thus the continuum approaches mentioned above do not apply directly. In the step-flow model, a stepped surface changes its morphology by lateral motion of steps [3, 33, 29, 24, 1]. For an unstrained film, a step can be viewed as a force dipole on the film surface. The elastic dipole interaction force decays as $1/r^3$ for two parallel straight steps with distance r. A step on the surface of a strained film (in heteroepitaxy) can be approximated by a force monopole. The elastic monopole interaction force decays as $1/r$ for two parallel straight steps with distance r. Expansions with higher order terms for the elastic effects of surface steps were obtained in [37, 44, 20].

The force dipole effect between steps on unstrained epitaxial surfaces is relatively well modeled in the frameworks of both the discrete step dynamics and the continuum theory [24, 21, 31, 35, 33, 16, 32, 29, 26, 27, 28, 25]. In the framework of the continuum theory, it is well known that the elastic effect on such a stepped surface is quite different from that on a surface above roughening transition temperature, due to the discrete structures on the stepped surface.

[*] Received by the editors January 5, 2008; accepted for publication (in revised form) November 12, 2008; published electronically February 19, 2009. This work was partially supported by Hong Kong RGC Competitive Earmarked Research grant 604604, UGC Targets of Opportunity AoE-MG/P05/06, and Croucher Foundation grant Z0138.

http://www.siam.org/journals/siap/69-5/71258.html

[†] Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (xuhy@ust.hk, maxiang@ust.hk).

Several models can be found in the literature on the long-range force monopole interaction on heteroepitaxial surfaces with straight and parallel steps (1+1 dimensional models). Alerhand et al. [1] studied the spontaneous formation of stress domains on crystal surfaces. Tersoff et al. [47, 23] described the step bunching phenomenon by a step dynamics model accounting for the attractive force monopole interaction due to misfit stress and the repulsive force dipole interaction. Duport et al. [5, 6] proposed models that take into account the force monopole and force dipole interactions as well as the elastic interaction between adatoms and steps and the effect of the Schwoebel barrier. Kaganer and Ploog [18] studied the energy of a strained stepped surface including the force dipole and force monopole effects of the steps and the interaction between them. Shenoy and Freund [39] obtained a continuum model based on continuum variational principles. Besides the force dipole and force monopole effects of the steps, they also showed that the dependence of the step line energy on the sign of the misfit strain may lead to nucleation of steps without energy barrier. Even though the discrete features of the step line energy and force dipole interaction between steps were included in these continuum models [18, 39], for the long-range elastic interaction due to misfit, the traditional expression above the roughening transition temperature was directly used. Xiang [49] and Xiang and E [51] derived a continuum model rigorously by taking the continuum limit of the discrete step dynamics models of Tersoff et al. [47, 23] and Duport et al. [5, 6]. For the misfit elastic effect, in addition to the widely used integral expression above the roughening transition temperature, there is another term incorporating the discrete features of the stepped surface in their continuum model. This additional term is crucial in modeling the step bunching instability on stepped surfaces [51].

A few models for crystal surfaces in $2 + 1$ dimensions accounting for the force monopole elastic effect of curved steps have also been proposed. Tersoff and Pehlke [46] analyzed step undulation instability of a stepped Si(001) surface which is subject to a force monopole effect at the steps, and their results agree well with the experimental observations obtained by Tromp and Reuter [48]. Houchmandzadeh and Misbah [15] studied the force dipole and force monopole elastic interactions between modulated steps. Kukta and Bhattacharya [19] proposed a $2 + 1$ step-flow model that accounts for both the elastic effects and terrace diffusion. Léonard and Tersoff [22] compared the different instability modes of a stepped surface under stress for both permeable and impermeable steps. Shenoy [38] studied the growth of epitaxial nanowires by controlled coarsening of strained islands. Haußer, Jabbour, and Voigt [12] proposed a step-flow model for the heteroepitaxial growth of strained, substitutional, binary alloy films with phase segregation. These models are all within the framework of discrete step dynamics. Within the continuum framework, Kaganer and Ploog [18] investigated the energetics of strained axially symmetric cone-shaped stepped surfaces; Ramasubramaniam and Shenoy [36] generalized the $1 + 1$ dimensional continuum model of Shenoy and Freund [39] to $2 + 1$ dimensions. However, continuum expressions above the roughening transition temperature were directly used in these continuum models. Still lacking is a continuum theory that accounts for the long-range elastic effect and the discrete features for a strained epitaxial film with a stepped surface in $2 + 1$ dimensions, as the continuum equation proposed in [49, 51] for a surface with straight steps in $1 + 1$ dimensions.

In this paper, we present such a continuum model for the long-range elastic interaction on the stepped surface of a strained film in $2 + 1$ dimensions. The continuum model is derived from the discrete model based on the BCF theory [3] that incorporates the processes of adatom diffusion on the terraces and attachment-detachment of adatoms to the steps [3, 33]. The continuum model is derived under the condition that the lattice constant and the average distance between adjacent steps are very small compared with the length unit of the continuum model, and thus there are many steps contained in a unit area of the domain of the continuum model. The derived continuum model contains new terms representing the contribution to the step line energy from this long-range elastic interaction, in addition to the traditional continuum expression for the surface above the roughening transition temperature. We focus on derivation in this paper. Validation of the model and simulation results using the model will be presented elsewhere [55].

The rest of this paper is organized as follows. In section 2, we briefly review the elastic effects on epitaxial surfaces and currently available continuum models for them. In section 3, we give the discrete model for the long-range elastic interaction on the stepped surface of a strained film. In section 4, we present the details of the derivation of the continuum model for the long-range elastic interaction on a stepped surface by taking the continuum limit from the discrete model. The results are summarized and discussion is made in section 5.

2. Currently available continuum models for epitaxial surfaces under elastic effects.

In this section, we briefly review the elastic effects on epitaxial surfaces and currently available continuum models for them. More details can be found in the books and reviews [33, 10, 9, 29] and other references in this section.

In heteroepitaxial growth, the misfit of the lattice constants is defined by

$$\varepsilon_0 = \frac{a_f - a}{a}, \tag{2.1}$$

where a_f and a are the lattice constants of the film and the substrate, respectively. This misfit results in strain and stress fields in the film and the substrate. For an isotropic film with flat surface and infinite substrate, there is a constant stress field in the bulk of the film. When the height of the film is in the z direction, the nonzero components of the stress tensor in the film are

$$\sigma_{xx} = \sigma_{yy} = \sigma_0 = \frac{2G(1+\nu)\varepsilon_0}{1-\nu}, \tag{2.2}$$

when the film and the substrate have the same shear modulus G and Poisson's ratio ν.

Above the roughening transition temperature, the surface can be modeled as a continuously changed profile without discrete structures on it; see Figure 2.1. Without the deposition flux, the surface morphological evolution due to surface diffusion is given by [30]

$$\frac{\partial h}{\partial t} = (1 + |\nabla h|^2)^{\frac{1}{2}} \nabla_s \cdot (D_s \nabla_s \mu), \tag{2.3}$$

where $h = h(x, y)$ is the height of the surface, ∇_s is the surface gradient operator, D_s is the surface diffusion coefficient, μ is the total chemical potential,

$$\mu = \mu_0 + \mu_m, \tag{2.4}$$

μ_0 is the chemical potential due to the surface energy, and μ_m is the chemical potential due to the elastic energy.

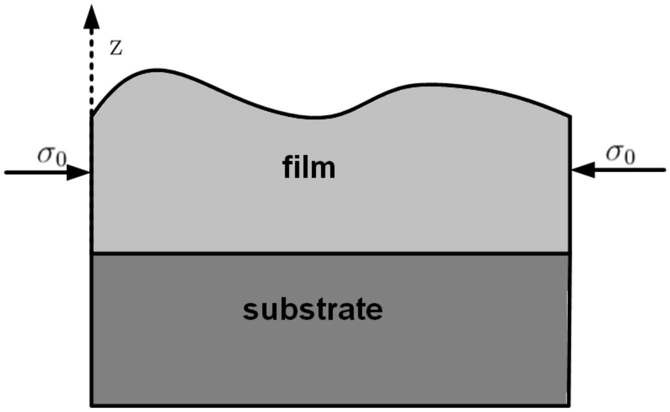


FIG. 2.1. *The film undergoes a stress due to the misfit in heteroepitaxy.*

The chemical potential due to the surface energy μ_0 is given by [30, 13]

$$\mu_0 = \gamma\kappa, \tag{2.5}$$

where γ is the surface energy density, $\kappa = \kappa_1 + \kappa_2$, κ_1 and κ_2 are the two principal values of the curvature of the surface.

The chemical potential due to the elastic energy μ_m is given by the elastic energy density on the surface [2, 11, 43]. For a nonflat surface, the chemical potential μ_m has to be obtained by solving an elasticity system in the film and the substrate. For a film with slightly modulated surface subject to the misfit stress, the elasticity system is approximately equivalent to that in a film with flat surface and subject to a traction [2, 11, 43, 42, 8, 33, 10, 9]

$$\mathbf{T} = \sigma_0(h_x, h_y, 0) \tag{2.6}$$

on its surface. In this case, μ_m is given by [17]

$$\mu_m(x,y) = -\frac{(1-\nu)\sigma_0^2}{2\pi G}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(x-\xi)h_x(\xi,\eta)+(y-\eta)h_y(\xi,\eta)}{[(x-\xi)^2+(y-\eta)^2]^{3/2}}\,d\xi d\eta, \tag{2.7}$$

when the surface $h(x,y)$ is defined in the whole two-dimensional space.

Below the roughening transition temperature, an epitaxial surface consists of atomic-height steps and atomic flat terraces; see Figure 2.2. The surface changes its morphology by motion, nucleation, and annihilation of steps. The evolution equation is given by [21, 31, 35, 16, 18, 49, 39, 51, 36, 38, 26, 27, 28, 25]

$$\frac{\partial h}{\partial t} = \nabla\cdot(D_0\nabla\mu), \tag{2.8}$$

where D_0 is the mobility. The mobility D_0 depends on the adatom diffusion on the terraces and attachment-detachment of adatoms along the steps. Derivations of this mobility from the BCF models [3] can be found for surfaces with straight steps in [49, 51], for conical surfaces with circular steps in [16, 18, 27], and for general stepped surfaces in $2+1$ dimensions in [25].

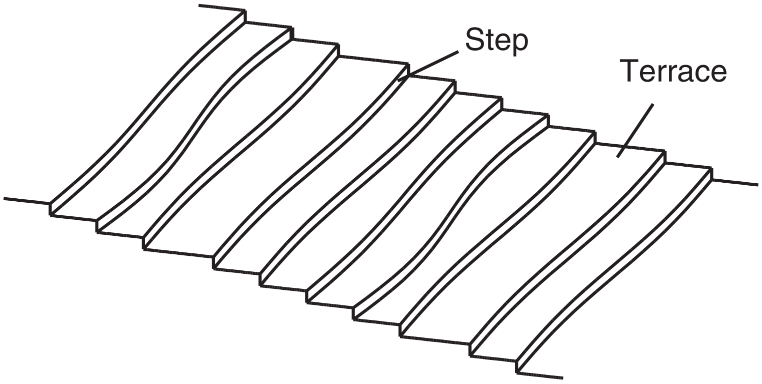


FIG. 2.2. *A stepped epitaxial surface.*

The total chemical potential μ on a misfit-strained stepped surface can be written as

$$\mu = \mu_s + \mu_d + \widehat{\mu}_m. \tag{2.9}$$

The first term μ_s is the chemical potential due to the step line energy

$$\mu_s = -\nabla \cdot \left(g_1 \frac{\nabla h}{|\nabla h|} \right) = \frac{\delta E_{\text{line}}}{\delta h}, \tag{2.10}$$

where

$$E_{\text{line}} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1 |\nabla h| dx dy \tag{2.11}$$

is the step line energy, and g_1 is the step line energy density. The second term μ_d is the chemical potential due to the interaction between steps in unstrained films which can be approximated by force dipole interaction

$$\mu_d = -\nabla \cdot (g_3 |\nabla h| \nabla h) = \frac{\delta E_{\text{dipole}}}{\delta h}, \tag{2.12}$$

where

$$E_{\text{dipole}} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{g_3}{3} |\nabla h|^3 dx dy \tag{2.13}$$

is the energy due to the force dipole interaction, and g_3 is the strength of this interaction.

The last term, $\widehat{\mu}_m$, in the above total chemical potential expression is the contribution of the long-range elastic interaction due to the misfit. The chemical potential for this long-range elastic interaction above the roughening transition temperature given in (2.7) was directly used for a surface with straight steps in [39], for a conical surface with circular steps in [18], and for a stepped surface in 2+1 dimensions in [36].

Xiang [49] and Xiang and E [51] derived a continuum model for epitaxial surfaces with straight steps rigorously by taking the continuum limit of the discrete models of steps [47, 23, 5, 6]. Their result for the force monopole and dipole elastic effects of steps is

$$\widehat{\mu}(x) = -\frac{(1-\nu)\sigma_0^2}{\pi G} \int_{-\infty}^{\infty} \frac{h'(\xi)}{x-\xi} d\xi - \frac{(1-\nu)\sigma_0^2 a}{2\pi G} \frac{h_{xx}(x)}{|h_x(x)|} \tag{2.14}$$

for a monotonic surface. The first term in (2.14) is the same as the chemical potential above the roughening transition temperature given in (2.7) in $1+1$ dimensions. The second term is the correction due to the discrete step-terrace structure of the surface. This continuum model of the chemical potential is also the variation of an elastic energy:

$$E_{\text{misfit}} = -\frac{(1-\nu)\sigma_0^2}{2\pi G}\int_{-\infty}^{\infty}\left(h(x)\int_{-\infty}^{\infty}\frac{h'(\xi)}{x-\xi}d\xi + a|h_x|\log|h_x|\right)dx. \tag{2.15}$$

In this paper, we shall generalize this work to epitaxial surfaces with curved steps in $2+1$ dimensions.

3. Discrete model for the elastic interaction between steps due to misfit.

In this section, we give the discrete model for the chemical potential of a stepped surface due to the misfit-induced long-range elastic interaction in $2+1$ dimensions. Following [1, 46, 5, 6, 15, 47, 23, 33, 19, 18, 22, 29, 38], this chemical potential can be obtained by considering the surface height $h(x,y)$ as a mathematical step function whose jump at a step is the lattice constant a in the continuum model (2.7). We assume that the steps $\{\gamma_j\}$, $j=\dots,-2,-1,0,1,2,\dots$, are smooth plane curves. Using the relation

$$\nabla h = -a\sum_j \delta(\gamma_j)\mathbf{n}_j \tag{3.1}$$

in (2.7), where $\delta(\gamma_j)$ is the one-dimensional Dirac delta function in the normal direction $\mathbf{n}_j$ of γ_j, the chemical potential at a point $\mathbf{X}$ on step γ_n is

$$\widehat{\mu}_m(\mathbf{X}) = \widehat{\mu}_m^{\text{int}}(\mathbf{X}) + \widehat{\mu}_m^{\text{self}}(\mathbf{X}), \tag{3.2}$$

with

$$\widehat{\mu}_m^{\text{int}}(\mathbf{X}) = \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\left(\sum_{j\neq n}\int_{\gamma_j}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_j(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl\right) \tag{3.3}$$

and

$$\widehat{\mu}_m^{\text{self}}(\mathbf{X}) = \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\left(\int_{-\infty}^{\infty}\delta(\omega)\,d\omega\int_{\gamma_\omega}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_\omega(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl\right), \tag{3.4}$$

where $\mathbf{X}_1$ is a point that varies along the curves in these line integrals, the curve $\gamma_\omega = \{\mathbf{X}_0 + \omega\mathbf{n}_n(\mathbf{X}_0) : \mathbf{X}_0 \in \gamma_n\}$, and $\mathbf{n}_\omega$ is its unit normal vector. Note that there are two possible directions for the normal vector of a step. In (3.1) and throughout this paper, we choose the normal direction of a step to be the direction in which the surface height h is decreasing.

The delta function $\delta(\omega)$ in the chemical potential due to the self-interaction $\widehat{\mu}_m^{\text{self}}(\mathbf{X})$ has to be regularized to avoid nonphysical singularity. (The same notation is used for simplicity.) The width of the regularization of $\delta(\omega)$, which reflects the detailed structure of the step and can be determined from atomistic calculations, is of the order of the lattice constant a. The regularization or cut-off is commonly used to remove nonphysical singularities in the models of steps [1, 19, 22, 29, 38] or other line defects such as dislocations [14]. We also assume that the regularized delta function $\delta(\omega)$ has compact support and $\delta(-\omega)=\delta(\omega)$.

Suppose that the step γ_n is parameterized by its arclength s, and the unit normal vector at a point $\mathbf{X}_1(s) \in \gamma_n$

$$\mathbf{n}_n(\mathbf{X}_1) = (-y'(s), x'(s)) \tag{3.5}$$

is in the direction in which the surface is decreasing. The curve γ_ω can also be parameterized by s (no longer arclength) as

$$(x_\omega(s)\,, y_\omega(s)) = (x(s) - \omega y'(s)\,, y(s) + \omega x'(s)), \tag{3.6}$$

and its tangent vector and normal vector (not normalized) are

$$(x'_\omega(s)\,, y'_\omega(s)) = (1 - \kappa(s)\omega)(x'(s)\,, y'(s)) \tag{3.7}$$

and

$$(-y'_\omega(s)\,, x'_\omega(s)) = (1 - \kappa(s)\omega)(-y'(s)\,, x'(s)), \tag{3.8}$$

respectively, where $\kappa(s)$ is the curvature of γ_n at point $\mathbf{X}_1(s)$. Note that $(x''(s)\,, y''(s)) = \kappa(s)(-y'(s)\,, x'(s))$. Thus we can rewrite the chemical potential due to the self-interaction in (3.4) as

$$\widehat{\mu}_m^{\text{self}}(\mathbf{X}) = \frac{(1-\nu)\sigma_0^2 a}{2\pi G} \int_{-\infty}^{\infty} \delta(\omega)\, d\omega \int_{\gamma_n} (1 - \kappa(\mathbf{X}_1)\omega) \frac{(\mathbf{X} - \mathbf{X}_1 - \omega \mathbf{n}_n(\mathbf{X}_1)) \cdot \mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X} - \mathbf{X}_1 - \omega \mathbf{n}_n(\mathbf{X}_1)\|^3}\, dl. \tag{3.9}$$

4. Derivation of the continuum model. Although (3.2)–(3.4) give an exact expression for the misfit-induced long-range elastic interaction on a stepped surface, in the framework of a continuum model, the stepped surface is often described by a smooth profile without resolving the details of the steps; see the surfaces $z = h(x, y)$ in Figure 4.1. In this section, we will derive a continuum model for the long-range elastic interaction on such a surface from the discrete model given by (3.2)–(3.4) by letting the lattice constant $a \to 0$. Similarly to the derivation for surfaces with straight steps in [49, 51], we will find the continuum approximation by considering the difference between this discrete model and the continuum expression on a surface without steps given by (2.7).

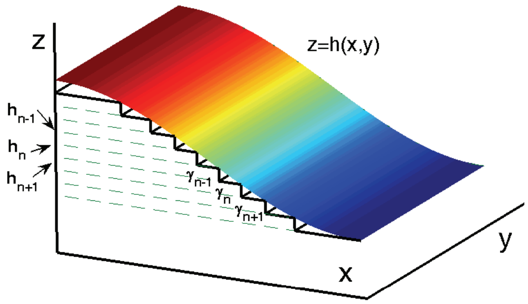


FIG. 4.1. *A stepped surface described by a smooth profile $z = h(x, y)$.*

4.1. Assumptions. We assume that in the length unit of the continuum model, the lattice constant $a \ll 1$, which means that there are many steps in a unit area of the domain of the continuum model, and the curvature radius of the steps is $O(1)$. The surface in the continuum model is a smooth profile $h(x, y)$, and a step γ_n with height h_n is described by the contour line $\gamma_n = \{(x, y)|h(x, y) = h_n\}$; see Figure 4.1. We assume that $h(x, y)$ has up to fourth order bounded partial derivatives, and $|\nabla h(x, y)| > c_0$, where $c_0 > 0$ is a constant. When the domain size is infinity, we assume that the integral expression in (2.7) converges, the line integrals in the discrete model in (3.3) and (3.4) converge absolutely, and the summation in the discrete model in (3.3) converges. Without loss of generality, the whole xy plane is used as the domain in our derivation.

4.2. Integral expression reformulated. In this subsection, we rewrite the integral expression given by (2.7) using parameter z and the parameter of the contour line

$$\gamma_z = \{(x, y)|h(x, y) = z\} \tag{4.1}$$

for the surface $z = h(x, y)$. From the assumptions of $h(x, y)$ given above, the contour line γ_z is smooth, with unit normal vector

$$\mathbf{n}_z(x, y) = -\frac{\nabla h(x, y)}{|\nabla h(x, y)|} \tag{4.2}$$

and bounded curvature

$$\kappa(x, y) = \nabla \cdot \left(\frac{\nabla h(x, y)}{|\nabla h(x, y)|} \right) \tag{4.3}$$

for $(x, y) \in \gamma_z$.

Using the new parameters, the integral expression (2.7) for the chemical potential at a point (x, y) on step γ_n, i.e., $h(x, y) = h_n$, can be rewritten as

$$\begin{aligned} \mu_m(x, y) &= -\frac{(1-\nu)\sigma_0^2}{2\pi G} \int_{-\infty}^{\infty} dz \int_{\gamma_z} \frac{(x-\xi)h_x(\xi, \eta) + (y-\eta)h_y(\xi, \eta)}{[(x-\xi)^2 + (y-\eta)^2]^{3/2}} \frac{1}{\sqrt{h_x^2(\xi, \eta) + h_y^2(\xi, \eta)}}\, dl \\ &= \frac{(1-\nu)\sigma_0^2}{2\pi G} \int_{-\infty}^{\infty} dz \int_{\gamma_z} \frac{(\mathbf{X} - \mathbf{X}_1) \cdot \mathbf{n}_z(\mathbf{X}_1)}{\rho^3}\, dl, \end{aligned} \tag{4.4}$$

where $\rho = \sqrt{(x-\xi)^2 + (y-\eta)^2}$, $\mathbf{X}_1 = (\xi, \eta)$ varies along γ_z, and $\mathbf{X} = (x, y)$.

Using the above formulation for the integral expression, the discrete formulation (3.2)–(3.4) can be roughly considered as a numerical discretization for the outer integral with respect to z of the integral expression. Since the inner integral in (4.4),

$$J(z) = \int_{\gamma_z} \frac{(\mathbf{X} - \mathbf{X}_1) \cdot \mathbf{n}_z(\mathbf{X}_1)}{\rho^3}\, dl, \tag{4.5}$$

is singular when $z = h_n = h(x, y)$, here we derive the asymptotic behavior of this singularity as $z \to h_n$, which will be used later to find the difference between the discrete and integral expressions accurately. The derivation method is similar to that for the asymptotic behavior of vortices in the Ginzburg–Landau equation [7, 34].

Suppose that $\mathbf{X}_0$ is the point on γ_z such that the point $\mathbf{X}$ lies on the normal direction (positive or negative) of γ_z at $\mathbf{X}_0$. Let $(\mathbf{t}_z, \mathbf{n}_z)$ be the Frenet coordinates at $\mathbf{X}_0$, where $\mathbf{t}_z$ is the unit tangent vector of γ_z, and $\mathbf{n}_z$ is the unit normal vector of γ_z defined in (4.2). Let s be a shifted arclength of γ_z such that $\mathbf{X}_0 = \gamma_z(0)$, and let $\kappa_z(s)$ be the curvature along γ_z. If $\kappa_z = \kappa_z(0)$ is the curvature of γ_z at $\mathbf{X}_0$ and $\kappa_z' = \kappa_z'(0)$, a point $\mathbf{X}_1$ on γ_z near $\mathbf{X}_0$ has the representation $\mathbf{X}_1 = \alpha(s)\mathbf{t}_z + \beta(s)\mathbf{n}_z$, where $\alpha(s) = s - \frac{1}{6}\kappa_z^2 s^3 + \cdots$ and $\beta(s) = \frac{1}{2}\kappa_z^2 s^2 + \frac{1}{6}\kappa_z' s^3 + \cdots$, and $\mathbf{X} = d\mathbf{n}_z$. Thus d is the signed distance from point $\mathbf{X}$ to the curve γ_z. We will find the asymptotic behavior of the integral $J(z)$ in (4.5) as $d \to 0$. Note that the signed distance d can be expressed in terms of z near h_n:

$$d = -\frac{z - h_n}{\sqrt{h_x^2(\mathbf{X}_0) + h_y^2(\mathbf{X}_0)}} + O((z - h_n)^2),\ z \to h_n, \tag{4.6}$$

whose proof can be found in Appendix A. (Note that the notation is different in Appendix A. The points $\mathbf{X}_0$ and $\mathbf{X}$ here are the points $\mathbf{X}$ and $\mathbf{P}$ in Appendix A, respectively. See Figure A.1.)

We first divide the integral in (4.5) into two parts: $J(z) = J_1(z) + J_2(z)$ with

$$J_1(z) = \int_{\gamma_z^1} \frac{(\mathbf{X} - \mathbf{X}_1) \cdot \mathbf{n}_z(\mathbf{X}_1)}{\rho^3}\, dl, \quad J_2(z) = \int_{\gamma_z^2} \frac{(\mathbf{X} - \mathbf{X}_1) \cdot \mathbf{n}_z(\mathbf{X}_1)}{\rho^3}\, dl, \tag{4.7}$$

where γ_z^1 is the segment of γ_z for $s \in [-1, 1]$, $\gamma_z^2 = \gamma_z - \gamma_z^1$, and $\mathbf{n}_z(\mathbf{X}_1) = -\beta'(s)\mathbf{t}_z + \alpha'(s)\mathbf{n}_z$. Using the expansions of $\alpha(s)$ and $\beta(s)$ and change of variable $s = \lambda|d|$ [34], we have

$$\begin{aligned}
J_1(z) &= \int_{-1}^{1} \frac{\alpha(s)\beta'(s) - \alpha'(s)\beta(s) + \alpha'(s)d}{[\alpha^2(s) + (\beta(s) - d)^2]^{3/2}}\, ds \\
&= \int_{-1}^{1} \frac{d + \frac{1}{2}\kappa_z s^2 + O(s^3 + s^2 d)}{(s^2 + d^2 - d\kappa_z s^2)^{3/2}[1 + O(s^2 + sd)]}\, ds \\
&= \int_{-1}^{1} \frac{d + \frac{1}{2}\kappa_z s^2 + O(s^2 d + s d^2)}{(s^2 + d^2 - d\kappa_z s^2)^{3/2}}\, ds + O(1) \\
&= \int_{-|d|^{-1}}^{|d|^{-1}} \frac{d|d| + \frac{1}{2}d^2|d|\kappa_z\lambda^2 + O(\lambda^2 d^4 + \lambda d^4)}{(\lambda^2 d^2 + d^2 - d^3\kappa_z\lambda^2)^{3/2}}\, d\lambda + O(1) \\
&= \int_{-|d|^{-1}}^{|d|^{-1}} \frac{1 + \frac{1}{2}d\kappa_z\lambda^2 + O(\lambda^2 d^2 + \lambda d^2)}{d(\lambda^2 + 1 - d\kappa_z\lambda^2)^{3/2}}\, d\lambda + O(1) \\
&= -\kappa_z \log|d| + \frac{2}{d} + O(1) \\
&= -\kappa \log|d| + \frac{2}{d} + O(1), \qquad (4.8)
\end{aligned}$$

as $d \to 0$, where κ is the curvature of the step γ_n at $\mathbf{X}$. Since $J_2(z) = O(1)$, we have

$$J(z) = -\kappa \log|d| + \frac{2}{d} + O(1),\ d \to 0. \tag{4.9}$$

This asymptotic behavior holds uniformly when $z \to h_n$ due to the smoothness assumption of the surface $h(x, y)$ and the convergence of the line integral.

Using (4.6) and (4.9), and $\mathbf{X}=\mathbf{X}_0+O(z-h_n)$, we have the asymptotic approximation of line integral in (4.5):

$$J(z)=-\kappa\log|z-h_n|-\frac{2\sqrt{h_x^2(\mathbf{X})+h_y^2(\mathbf{X})}}{z-h_n}+O(1),\ z\to h_n. \tag{4.10}$$

4.3. Difference between the discrete model and the continuum expression. When the continuum expression $\mu_m(x,y)$ is written in the form in (4.4), the discrete model $\widehat{\mu}_m(x,y)$ given by (3.2)–(3.4) can be regarded as a numerical scheme of this continuum expression. Thus, the continuum expression with the leading order error terms of this numerical scheme will give an accurate continuum approximation of the discrete model. The error of this numerical scheme can be found using the following theorem, whose proof is given in Appendix B.

THEOREM 4.1. *Suppose that interval $[a,b]$ is divided into m subintervals with $\Delta x=(b-a)/m$, $x_j=a+(j-1)\Delta x$, $j=1,\dots,m+1$. Let $G(x)=g_1(x)\log|x-t|+g_2(x)/(x-t)+g_3(x)$ with $t=x_{j_0}$ for some j_0, where $g_1(x)$, $g_2(x)$, and $g_3(x)$ are twice continuously differentiable functions. Then*

$$\begin{aligned}\int_a^b G(x)\,dx&=\Delta x\left(\frac{G(a)+G(b)}{2}+\sum_{2\le j\le m,j\ne j_0}G(x_j)\right)\\&\quad+\int_{t-\frac{\Delta x}{2}}^{t+\frac{\Delta x}{2}}G(x)\,dx-(\log\pi-1)g_1(t)\Delta x+O(\Delta x^2).\end{aligned} \tag{4.11}$$

Using this theorem for $G(z)=\frac{(1-\nu)\sigma_0^2}{2\pi G}J(z)$, where $J(z)$ is given by (4.5) with the asymptotic behavior as $z\to h_n$ given by (4.10), and the discretization $z=h_j$, $j=\dots,-1,0,1,2,\dots$, we have, as $a\to 0$,

$$\begin{aligned}\mu_m(x,y)&=\widehat{\mu}_m^{\text{int}}(x,y)\\&\quad+\frac{(1-\nu)\sigma_0^2}{2\pi G}\left(\int_{h_n-\frac{a}{2}}^{h_n+\frac{a}{2}}\left(\int_{\gamma_z}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_z(\mathbf{X}_1)}{\rho^3}dl\right)dz+(\log\pi-1)\kappa a\right)\\&\quad+O(a^2),\end{aligned} \tag{4.12}$$

or

$$\begin{aligned}\widehat{\mu}_m(x,y)&=\widehat{\mu}_m^{\text{int}}(x,y)+\widehat{\mu}_m^{\text{self}}(x,y)\\&=\mu_m(x,y)+\widehat{\mu}_m^{\text{self}}(x,y)-\mu_m^{\text{self}}(x,y)-\frac{(1-\nu)\sigma_0^2}{2\pi G}(\log\pi-1)\kappa a+O(a^2),\end{aligned} \tag{4.13}$$

where

$$\mu_m^{\text{self}}(x,y)=\frac{(1-\nu)\sigma_0^2}{2\pi G}\int_{h_n-\frac{a}{2}}^{h_n+\frac{a}{2}}\left(\int_{\gamma_z}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_z(\mathbf{X}_1)}{\rho^3}dl\right)dz. \tag{4.14}$$

Now, to find the continuum approximation of $\widehat{\mu}_m(x,y)$, it remains to find the difference between the self-interaction in the discrete model $\widehat{\mu}_m^{\text{self}}(x,y)$, given by (3.4), and $\mu_m^{\text{self}}(x,y)$, given by (4.14), which can be explained as the self-interaction in the continuum model.

In order to find the difference between $\widehat{\mu}_m^{\text{self}}(x,y)$ and $\mu_m^{\text{self}}(x,y)$, similar to $\widehat{\mu}_m^{\text{self}}(x,y)$ given by (3.9), we also rewrite the inner integral in $\mu_m^{\text{self}}(x,y)$ as an integral along γ_n.

When the lattice constant a is very small compared with the minimum radius of curvature of contour lines of h, which exists due to the uniform boundedness of the curvature, the contour line γ_z when $z \in [h_n - \frac{a}{2}, h_n + \frac{a}{2}]$ can be written as

$$\gamma_z = \{\mathbf{X}_1 + d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1) : \mathbf{X}_1 \in \gamma_n\}, \tag{4.15}$$

where $\omega = z - h_n$, $\mathbf{n}_n(\mathbf{X}_1)$ is the unit normal vector of γ_n at the point $\mathbf{X}_1$ on γ_n, and $d(\mathbf{X}_1, \omega)$ is the signed distance between a point $\mathbf{X}_1$ on γ_n and a point $\mathbf{P}$ on the nearby contour line γ_z such that $\mathbf{PX}_1$ is parallel to the normal direction of γ_n at $\mathbf{X}_1$.

Suppose that γ_n is parameterized by its arclength s: $\mathbf{X}_1 = (x(s), y(s)) \in \gamma_n$, with unit tangent vector $\mathbf{t}_n(\mathbf{X}_1) = (x'(s), y'(s))$ and unit normal vector $\mathbf{n}_n(\mathbf{X}_1) = (-y'(s), x'(s))$. Then γ_z can also be parameterized by s (no longer arclength) as

$$(\widehat{x}_\omega(s), \widehat{y}_\omega(s)) = (x(s) - d(s, \omega)y'(s)\,, y(s) + d(s, \omega)x'(s)), \tag{4.16}$$

where $d(s, \omega) = d(\mathbf{X}_1, \omega)|_{\mathbf{X}_1=(x(s),y(s))}$ is the signed distance from the point $\mathbf{P}$ to the point $\mathbf{X}_1$ which is positive in the direction of $\mathbf{n}_n(\mathbf{X}_1)$. The tangent and normal vectors of γ_z (not normalized) are

$$(\widehat{x}'_\omega(s), \widehat{y}'_\omega(s)) = (1 - \kappa(s)d(s, \omega))(x'(s)\,, y'(s)) + d_s(s, \omega)(-y'(s)\,, x'(s)) \tag{4.17}$$

and

$$(-\widehat{y}'_\omega(s), \widehat{x}'_\omega(s)) = (1 - \kappa(s)d(s, \omega))(-y'(s)\,, x'(s)) - d_s(s, \omega)(x'(s)\,, y'(s)), \tag{4.18}$$

respectively, where $\kappa(s)$ is curvature of γ_n at $(x(s), y(s))$, and $d_s(s, \omega) = \frac{\partial}{\partial s} d(s, \omega)$. Hence, $\mu_m^{\text{self}}(x, y)$ defined in (4.14) can be written as

$$\mu_m^{\text{self}}(x, y) = \frac{(1 - \nu)\sigma_0^2}{2\pi G} \int_{-\frac{a}{2}}^{\frac{a}{2}} \left(\int_{\gamma_n} \frac{(\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)) \cdot \widetilde{\mathbf{n}_n}(\mathbf{X}_1)}{\|\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)\|^3} dl \right) d\omega, \tag{4.19}$$

where

$$\widetilde{\mathbf{n}}_n(\mathbf{X}_1) = (1 - \kappa(\mathbf{X}_1)d(\mathbf{X}_1, \omega))\mathbf{n}_n(\mathbf{X}_1) - d_s(\mathbf{X}_1, \omega)\mathbf{t}_n(\mathbf{X}_1). \tag{4.20}$$

4.4. Difference between $\widehat{\mu}_m^{\text{self}}(x, y)$ and $\mu_m^{\text{self}}(x, y)$. Now we want to find the asymptotic approximation, up to $O(a)$, of the difference between $\widehat{\mu}_m^{\text{self}}(x, y)$ given by (3.9) and $\mu_m^{\text{self}}(x, y)$ given by (4.19). We use a shifted arclength parameter s for γ_n with $\mathbf{X} = \gamma_n(0)$. When $\mathbf{X}_1 \in \gamma_n$ is close to $\mathbf{X}$,

$$\mathbf{X}_1 = \mathbf{X} + \left(s - \frac{\kappa^2 s^3}{6} \right) \mathbf{t} + \left(\frac{\kappa s^2}{2} + \frac{\kappa' s^3}{6} \right) \mathbf{n} + O(s^4), \tag{4.21}$$

where $\mathbf{t}$ and $\mathbf{n}$ are unit tangent and normal vectors of γ_n at $\mathbf{X}$, respectively, $\kappa = \kappa(s)|_{s=0}$ is the curvature of γ_n at $\mathbf{X}$, and $\kappa' = \frac{d\kappa(s)}{ds}|_{s=0}$.

Let γ_n^A be the segment of γ_n with $s \in [-a^{\frac{1}{4}}, a^{\frac{1}{4}}]$ and $\gamma_n^B = \gamma_n - \gamma_n^A$. Due to the smoothness assumptions of h, we have

$$\|\mathbf{X} - \mathbf{X}_1\|^2 \geq \frac{1}{2} a^{\frac{1}{2}}, \quad \mathbf{X}_1 \in \gamma_n^B, \tag{4.22}$$

when a is small enough compared with the minimum radius of the curvature of γ_n. Thus when $\omega = O(a)$, for $\mathbf{X}_1 \in \gamma_n^B$, we have

$$\begin{aligned} &\|\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1)\|^3 \\ &= (\|\mathbf{X}-\mathbf{X}_1\|^2 - 2\,\omega(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)+\omega^2)^{3/2} \\ &= \|\mathbf{X}-\mathbf{X}_1\|^3\left(1-2\,\omega\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^2}+\frac{\omega^2}{\|\mathbf{X}-\mathbf{X}_1\|^2}\right)^{3/2} \\ &= \|\mathbf{X}-\mathbf{X}_1\|^3(1+O(a^{\frac{3}{4}})). \end{aligned} \tag{4.23}$$

Using (4.22) and (4.23), when $\omega = O(a)$, we have

$$\int_{\gamma_n^B}(1-\kappa(\mathbf{X}_1)\omega)\frac{(\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1))\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1)\|^3}\,dl = \int_{\gamma_n^B}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl + O(a^{\frac{1}{4}}), \tag{4.24}$$

which implies that when the inner line integral is along γ_n^B in $\widehat{\mu}_m^{\mathrm{self}}(x,y)$ given by (3.9), we have

$$\begin{aligned} &a\int_{-\infty}^{\infty}\delta(\omega)d\omega\int_{\gamma_n^B}(1-\kappa(\mathbf{X}_1)\omega)\frac{(\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1))\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1)\|^3}\,dl \\ &= a\int_{\gamma_n^B}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl + O(a^{\frac{5}{4}}), \end{aligned} \tag{4.25}$$

when the delta function $\delta(\omega)$ has compact support with a width of $O(a)$.

Now we consider $\mu_m^{\mathrm{self}}(x,y)$ given by (4.19) when the inner line integral is along γ_n^B. First, it is shown in Appendix A that when ω is small,

$$d(\mathbf{X}_1,\omega) = \left. -\frac{\omega}{(h_x^2+h_y^2)^{1/2}} + \frac{h_x^2h_{xx}+2h_xh_yh_{xy}+h_y^2h_{yy}}{2\,(h_x^2+h_y^2)^{5/2}}\omega^2\right|_{\mathbf{X}_1=(x(s),y(s))} + O(\omega^3). \tag{4.26}$$

Thus we have

$$d_s(\mathbf{X}_1,\omega) = \left.\frac{h_x(x'h_{xx}+y'h_{xy})+h_y(x'h_{xy}+y'h_{yy})}{(h_x^2+h_y^2)^{3/2}}\omega\right|_{\mathbf{X}_1=(x(s),y(s))} + O(\omega^2) \tag{4.27}$$

and

$$\|\widetilde{\mathbf{n}}_n(\mathbf{X}_1)\| = 1 + O(\omega), \tag{4.28}$$

where $\widetilde{\mathbf{n}}_n(\mathbf{X}_1)$ is given by (4.20).

Using (4.22) and (4.26), when $\omega = O(a)$, for $\mathbf{X}_1 \in \gamma_n^B$, we have

$$\|\mathbf{X}-\mathbf{X}_1-d(\mathbf{X}_1,\omega)\mathbf{n}_n(\mathbf{X}_1)\|^3 = \|\mathbf{X}-\mathbf{X}_1\|^3(1+O(a^{\frac{3}{4}})). \tag{4.29}$$

Thus when $\omega = O(a)$, for $\mathbf{X}_1 \in \gamma_n^B$, using (4.28) and (4.29), we have

$$\int_{\gamma_n^B}\frac{(\mathbf{X}-\mathbf{X}_1-d(\mathbf{X}_1,\omega)\mathbf{n}_n(\mathbf{X}_1))\cdot\widetilde{\mathbf{n}_n}(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-d(\mathbf{X}_1,\omega)\mathbf{n}_n(\mathbf{X}_1)\|^3}\,dl = \int_{\gamma_n^B}\frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl + O(a^{\frac{1}{4}}), \tag{4.30}$$

which implies

$$\int_{-\frac{a}{2}}^{\frac{a}{2}} d\omega \int_{\gamma_n^B} \frac{(\mathbf{X}-\mathbf{X}_1-d(\mathbf{X}_1,\omega)\mathbf{n}_n(\mathbf{X}_1))\cdot\widetilde{\mathbf{n}_n}(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-d(\mathbf{X}_1,\omega)\mathbf{n}_n(\mathbf{X}_1)\|^3}\,dl$$
$$= a\int_{\gamma_n^B} \frac{(\mathbf{X}-\mathbf{X}_1)\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1\|^3}\,dl + O(a^{\frac{5}{4}}). \tag{4.31}$$

Next, we consider $\widehat{\mu}_m^{\rm self}(x,y)$ in (3.9) and $\mu_m^{\rm self}(x,y)$ in (4.19) when the inner line integrals are along γ_n^A. When $\mathbf{X}_1 \in \gamma_n^A$, $\mathbf{X}_1$ can be written as (4.21), and the unit tangent vector, unit normal vector, and curvature at $\mathbf{X}_1$ are

$$\mathbf{t}_n(\mathbf{X}_1) = \left(1-\frac{\kappa^2 s^2}{2}\right)\mathbf{t} + \left(\kappa s + \frac{\kappa' s^2}{2}\right)\mathbf{n} + O(s^3), \tag{4.32}$$

$$\mathbf{n}_n(\mathbf{X}_1) = \left(-\kappa s - \frac{\kappa' s^2}{2}\right)\mathbf{t} + \left(1-\frac{\kappa^2 s^2}{2}\right)\mathbf{n} + O(s^3), \tag{4.33}$$

and

$$\kappa(\mathbf{X}_1) = \kappa + O(s), \tag{4.34}$$

respectively, for $s \in [-a^{\frac{1}{4}}, a^{\frac{1}{4}}]$.

When $s \in [-a^{\frac{1}{4}}, a^{\frac{1}{4}}]$, $\omega = O(a)$, and a is small enough, using (4.21), (4.32)–(4.34), the integrand in (3.9) can be written as

$$(1-\kappa(\mathbf{X}_1)\omega)\frac{(\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1))\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1)\|^3}$$
$$= \frac{\kappa s^2/2-\omega+\kappa\omega^2+O(s^3+\omega s^2+\omega^2 s)}{[s^2+\omega^2-\kappa\omega s^2+O(s^4+\omega s^3)]^{3/2}}$$
$$= \frac{\kappa s^2/2-\omega+\kappa\omega^2}{(s^2+\omega^2)^{3/2}} - \frac{\frac{3}{2}\kappa\omega^2 s^2}{(s^2+\omega^2)^{5/2}} + O(1). \tag{4.35}$$

Hence, when $\mathbf{X}_1 \in \gamma_n^A$, the integral in $\widehat{\mu}_m^{\rm self}(x,y)$ given by (3.9) can be written as

$$\frac{(1-\nu)\sigma_0^2 a}{2\pi G}\int_{-\infty}^{\infty}\delta(\omega)\,d\omega\int_{\gamma_n^A}(1-\kappa(\mathbf{X}_1)\omega)\frac{(\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1))\cdot\mathbf{n}_n(\mathbf{X}_1)}{\|\mathbf{X}-\mathbf{X}_1-\omega\mathbf{n}_n(\mathbf{X}_1)\|^3}dl \tag{4.36}$$
$$= \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\int_{-\infty}^{\infty}\delta(\omega)d\omega\int_{-a^{\frac{1}{4}}}^{a^{\frac{1}{4}}}\left[\frac{\kappa s^2/2-\omega+\kappa\omega^2}{(s^2+\omega^2)^{3/2}} - \frac{\frac{3}{2}\kappa\omega^2 s^2}{(s^2+\omega^2)^{5/2}}\right]ds + O(a^{\frac{5}{4}})$$
$$= \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\int_{-\infty}^{\infty}\kappa\delta(\omega)(\log a^{\frac{1}{4}} + \log 2 - \log|\omega|)d\omega + O(a^{\frac{5}{4}})$$
$$= \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\kappa(\log 2a^{\frac{1}{4}} - \log r_c) + O(a^{\frac{5}{4}}),$$

when the regularized delta function $\delta(\omega)$ has compact support with a width of $O(a)$ and $\delta(-\omega) = \delta(\omega)$, where r_c is a parameter depending on the core of the step represented by the regularized delta function $\delta(\omega)$:

$$\log r_c = \int_{-\infty}^{\infty}\delta(\omega)\log|\omega|\,d\omega. \tag{4.37}$$

The inner integral with respect to s in (4.36) is calculated using change of variable $s = |\omega|\lambda$.

Now we consider $\mu_m^{\text{self}}(x, y)$ given by (4.19) when the inner line integral is along γ_n^A. First, from (4.26) and (4.27), we have

$$d(\mathbf{X}_1, \omega) = \beta_1\omega + \beta_2\omega^2 + \beta_3\omega s + O(\omega(s^2 + \omega^2)), \tag{4.38}$$

$$d_s(\mathbf{X}_1, \omega) = \beta_3\omega + O(\omega^2 + \omega s), \tag{4.39}$$

where

$$\beta_1 = -\frac{1}{(h_x^2 + h_y^2)^{1/2}}, \tag{4.40}$$

$$\beta_2 = \frac{h_x^2 h_{xx} + 2h_x h_y h_{xy} + h_y^2 h_{yy}}{2\,(h_x^2 + h_y^2)^{5/2}}, \tag{4.41}$$

$$\beta_3 = \frac{h_x(x' h_{xx} + y' h_{xy}) + h_y(x' h_{xy} + y' h_{yy})}{(h_x^2 + h_y^2)^{3/2}}, \tag{4.42}$$

in which all partial derivatives are values at $\mathbf{X} = (x, y)$.

Hence, when $s \in [-a^{\frac{1}{4}}, a^{\frac{1}{4}}]$, $\omega = O(a)$, and a is small enough, using (4.20), (4.21), (4.32)–(4.34), (4.38), and (4.39), the integrand in (4.19) can be written as

$$\begin{aligned}
&\frac{(\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)) \cdot \widetilde{\mathbf{n}}_n(\mathbf{X}_1)}{\|\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)\|^3} \\
&= \frac{-\beta_1\omega + (\kappa\beta_1^2 - \beta_2)\omega^2 + \kappa s^2/2}{(s^2 + \beta_1^2\omega^2)^{3/2}} + \frac{\frac{3}{2}\beta_1\omega(-\beta_1\kappa\omega s^2 + 2\beta_1\beta_3\omega^2 s + 2\beta_1\beta_2\omega^3)}{(s^2 + \beta_1^2\omega^2)^{5/2}} + O(1).
\end{aligned} \tag{4.43}$$

Using this expansion and change of variable $s = |\beta_1\omega|\lambda$, when $\mathbf{X}_1 \in \gamma_n^A$, the integral in $\mu_m^{\text{self}}(x, y)$ given by (4.19) can be written as

$$\begin{aligned}
&\frac{(1-\nu)\sigma_0^2}{2\pi G} \int_{-\frac{a}{2}}^{\frac{a}{2}} d\omega \int_{\gamma_n^A} \frac{(\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)) \cdot \widetilde{\mathbf{n}_n}(\mathbf{X}_1)}{\|\mathbf{X} - \mathbf{X}_1 - d(\mathbf{X}_1, \omega)\mathbf{n}_n(\mathbf{X}_1)\|^3} dl \\
&= \frac{(1-\nu)\sigma_0^2}{2\pi G} \int_{-\frac{a}{2}}^{\frac{a}{2}} d\omega \int_{-a^{\frac{1}{4}}}^{a^{\frac{1}{4}}} \left[\frac{-\beta_1\omega + (\kappa\beta_1^2 - \beta_2)\omega^2 + \kappa s^2/2}{(s^2 + \beta_1^2\omega^2)^{3/2}} \right. \\
&\qquad \left. + \frac{\frac{3}{2}\beta_1\omega(-\beta_1\kappa\omega s^2 + 2\beta_1\beta_3\omega^2 s + 2\beta_1\beta_2\omega^3)}{(s^2 + \beta_1^2\omega^2)^{5/2}} \right] ds + O(a^{\frac{5}{4}}) \\
&= \frac{(1-\nu)\sigma_0^2 a}{2\pi G} \left[\kappa \left(\log 2a^{\frac{1}{4}} + 1 + \log \frac{2\sqrt{h_x^2 + h_y^2}}{a} \right) \right. \\
&\qquad \left. + \frac{h_x^2 h_{xx} + 2h_x h_y h_{xy} + h_y^2 h_{yy}}{(h_x^2 + h_y^2)^{3/2}} \right] + O(a^{\frac{5}{4}}).
\end{aligned} \tag{4.44}$$

Using (4.25), (4.31), (4.36), and (4.44), we have

$$\begin{aligned}
&\widehat{\mu}_m^{\text{self}}(x, y) - \mu_m^{\text{self}}(x, y) \\
&= \frac{(1-\nu)\sigma_0^2 a}{2\pi G} \left[\kappa \left(-1 + \log \frac{a}{2r_c\sqrt{h_x^2 + h_y^2}} \right) - \frac{h_x^2 h_{xx} + 2h_x h_y h_{xy} + h_y^2 h_{yy}}{(h_x^2 + h_y^2)^{3/2}} \right] + O(a^{\frac{5}{4}}).
\end{aligned} \tag{4.45}$$

4.5. The continuum model for a stepped surface. Using (2.7), (4.3), (4.13), and (4.45), we have the continuum model for the long-range elastic interaction on a stepped surface in heteroepitaxy:

(4.46)

$$\begin{aligned}
&\widehat{\mu}_m(x,y)\\
&= \mu_m(x,y) + \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\left(\kappa\log\frac{a}{2\pi r_c\sqrt{h_x^2+h_y^2}} - \frac{h_x^2h_{xx}+2h_xh_yh_{xy}+h_y^2h_{yy}}{(h_x^2+h_y^2)^{3/2}}\right)\\
&\quad + O(a^{\frac{5}{4}})\\
&\approx -\frac{(1-\nu)\sigma_0^2}{2\pi G}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(x-\xi)h_x(\xi,\eta)+(y-\eta)h_y(\xi,\eta)}{[(x-\xi)^2+(y-\eta)^2]^{3/2}}d\xi d\eta\\
&\quad -\frac{(1-\nu)\sigma_0^2 a}{2\pi G}\left[\nabla\cdot\left(\frac{\nabla h}{|\nabla h|}\right)\log\frac{2\pi r_c|\nabla h|}{a} + \frac{\nabla h\, D^2h\,\nabla^T h}{|\nabla h|^3}\right],
\end{aligned}$$

where D^2h is the Hessian matrix of $h(x,y)$, and $\nabla h\, D^2 h\,\nabla^T h = h_x^2h_{xx}+2h_xh_yh_{xy}+h_y^2h_{yy}$. In this continuum expression, the integral term over the whole domain of the surface is the same as the expression of the misfit-induced elastic interaction above the roughening transition temperature, and the two additional terms incorporate the discrete feature of the stepped surface. It is easy to verify that this continuum model is reduced to the $1+1$ dimensional model for a surface with straight steps given by (2.14) [49, 51].

In particular, for an axisymmetric conical mound-like stepped surface $h(r)$ with $h'(r)<0$, where r is the radial coordinate, the result becomes

$$\begin{aligned}
\widehat{\mu_m}(r) &= \frac{(1-\nu)\sigma_0^2}{\pi G}\int_0^{\infty}\left(\frac{E(m)}{\hat{r}-r}+\frac{K(m)}{\hat{r}+r}\right)h'(\hat{r})\,d\hat{r}\\
&\quad + \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\left(\frac{1}{r}\log\frac{2\pi r_c|h'(r)|}{a}+\frac{h''(r)}{h'(r)}\right)+O(a^2), \qquad (4.47)
\end{aligned}$$

where $K(m) = \int_0^{\frac{\pi}{2}}\frac{d\theta}{\sqrt{1-m\sin^2\theta}}$ and $E(m) = \int_0^{\frac{\pi}{2}}\sqrt{1-m\sin^2\theta}\,d\theta$ are the complete elliptic integrals of the first and second kind, respectively, with $m = \frac{4\hat{r}r}{(\hat{r}+r)^2}$.

The derived continuum expression of chemical potential given by (4.46) can be written as the variation of an elastic energy

$$\widehat{\mu}_m(x,y) = \frac{\delta E_{\text{misfit}}}{\delta h}, \qquad (4.48)$$

where

$$\begin{aligned}
E_{\text{misfit}} = &-\frac{(1-\nu)\sigma_0^2}{4\pi G}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)\\
&\cdot\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(x-\xi)h_x(\xi,\eta)+(y-\eta)h_y(\xi,\eta)}{[(x-\xi)^2+(y-\eta)^2]^{3/2}}d\xi d\eta\right]dxdy\\
&+\frac{(1-\nu)\sigma_0^2 a}{2\pi G}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\nabla h|\log\frac{2\pi r_c|\nabla h|}{ea}dxdy. \qquad (4.49)
\end{aligned}$$

The first term in this expression is the traditional expression of the misfit elastic energy above the roughening transition temperature on surfaces with small amplitude modulation. The second term is new, which is the contribution to the step line energy from the force monopole interaction, and which gives the two local terms in the chemical potential in (4.46).

It is easy to verify that this misfit elastic energy is reduced to that in the $1+1$ dimensional continuum model for surfaces with straight steps given by (2.15) [49, 51]. Note that in the $1+1$ model, the term of $|h_x|$ with constant coefficient in the total energy, i.e., the step line energy with constant density, plays no role in the surface morphology evolution when the surface is monotonic; thus this kind of term does not appear in the $1+1$ model in (2.15).

The total misfit elastic energy of a surface consisting of a uniform step train has also been calculated using the discrete model [1], whose density is

$$e_{\text{misfit}} = -\frac{(1-\nu)\sigma_0^2 a^2}{2\pi G l}\log\frac{l}{\pi r_0}, \tag{4.50}$$

where l is the distance between two adjacent steps, and r_0 is a cut-off distance. In this case, using the relation $|h'(x)| = a/l$, our model (4.49) gives the misfit elastic energy density

$$e_{\text{misfit}} = -\frac{(1-\nu)\sigma_0^2 a^2}{2\pi G l}\log\frac{el}{2\pi r_c}, \tag{4.51}$$

which agrees with the result of the discrete model. (Note that the values of the cut-off distance are different in these two models.)

Finally, using (2.8), (2.9), (2.10), (2.12), and (4.46), we have the morphological evolution equation of a stepped surface:

$$\begin{aligned}\frac{\partial h}{\partial t} = \nabla\cdot\Bigg\{ D_0\nabla\Bigg[& -\nabla\cdot\left(g_1\frac{\nabla h}{|\nabla h|} + g_3|\nabla h|\nabla h\right) \\ & -\frac{(1-\nu)\sigma_0^2}{2\pi G}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(x-\xi)h_x(\xi,\eta)+(y-\eta)h_y(\xi,\eta)}{[(x-\xi)^2+(y-\eta)^2]^{3/2}}d\xi d\eta \\ & -\frac{(1-\nu)\sigma_0^2 a}{2\pi G}\nabla\cdot\left(\frac{\nabla h}{|\nabla h|}\right)\log\left(\frac{2\pi r_c|\nabla h|}{a}\right) - \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\frac{\nabla h\, D^2 h\,\nabla^T h}{|\nabla h|^3}\Bigg]\Bigg\}.\end{aligned} \tag{4.52}$$

The first two terms in this equation containing parameters g_1 and g_3 are due to the step line energy without misfit and the force dipole interaction between steps, respectively. The last three terms come from the long-range force monopole interaction due to the misfit stress.

For the special case of an axisymmetric conical stepped surface $h(r)$ with $h'(r) < 0$, the equation can be written as

$$\begin{aligned}\frac{\partial h}{\partial t} = \nabla\cdot\Bigg\{ D_0\nabla\Bigg[& \frac{g_1}{r} + \frac{g_3 {h'}^2}{r} + 2g_3 h'h'' + \frac{(1-\nu)\sigma_0^2}{\pi G}\int_0^{\infty}\left(\frac{E(m)}{\hat{r}-r} + \frac{K(m)}{\hat{r}+r}\right)h'(\hat{r})\,d\hat{r} \\ & + \frac{(1-\nu)\sigma_0^2 a}{2\pi G\, r}\log\frac{2\pi r_c|h'|}{a} + \frac{(1-\nu)\sigma_0^2 a}{2\pi G}\,\frac{h''}{h'}\Bigg]\Bigg\}.\end{aligned} \tag{4.53}$$

5. Conclusion and discussion. In this paper, we have presented a continuum model for the long-range elastic interaction on a stepped heteroepitaxial surface in $2+1$ dimensions. The continuum model is derived rigorously by taking the continuum limit from the discrete model for the interaction between steps. Compared with the integral expression above the roughening transition temperature, our model has additional terms that incorporate the discrete features of the stepped surface.

Our continuum model has a singularity when $|\nabla h(x,y)|=0$, which happens at a place where there is no step such as the top of a mound or a valley. This is consistent with our assumption in the derivation that $|\nabla h(x,y)|>c_0$, where $c_0>0$ is a constant. In fact, any continuum model containing step line energy becomes singular when $|\nabla h(x,y)|=0$ [21, 31, 35, 16, 39, 36, 26, 27, 28, 25] and may not describe the physics accurately at such places [35]. A continuum model that addresses this problem was given in [28]. A numerical treatment for this singularity can be found in [45, 39, 36].

Our continuum model is reduced to the $1+1$ dimensional continuum model in [49, 51] for a surface with straight steps. The energy for a surface with straight steps obtained using our model, which comes entirely from the two new terms, agrees with the result using the discrete model [1], which indicates the importance of these terms. Our model can be used to study the morphological instabilities of stepped surfaces under elastic effects. Such results, including linear instability analysis and numerical simulations in the nonlinear regime, as well as comparisons of our results with those of the discrete model, will be reported in a forthcoming paper [55].

Future work may also include the incorporation of the anisotropic mobility [25], the stress-dependent step line energy [39], the step line energy due to the interaction between the force monopole and dipole effects [18], or multispecies epitaxy [12].

Appendix A. Distance in the normal direction of a contour line of the surface *h*. Suppose that $\mathbf{X}=(x,y)$ is a point on the contour line of the surface h: $\gamma=\{(\xi,\eta):h(\xi,\eta)=h_n\}$. Let $\mathbf{P}=(u,v)$ be a point such that $\mathbf{XP}$ is parallel to the normal direction of γ at $\mathbf{X}$ and $h(u,v)=z$; see Figure A.1. We want to find the signed distance between points $\mathbf{P}$ and $\mathbf{X}$, which is positive in the direction of $-\nabla h(x,y)$, when z is close to h_n.

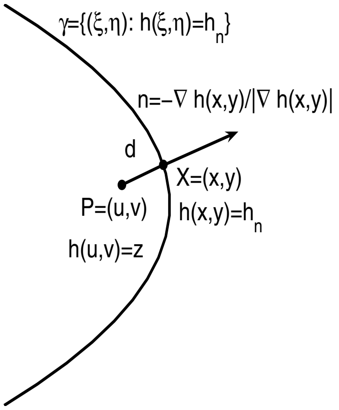


FIG. A.1. *Distance in the xy plane in the normal direction of a contour line of the surface h.*

Considering (u,v) as a function of z, we have

$$h(u(z),v(z))=z. \tag{A.1}$$

Since $\mathbf{XP}$ is parallel to the normal direction of γ at $\mathbf{X}$, we have

$$\frac{u(z)-x}{v(z)-y}=\frac{h_x^0}{h_y^0}. \tag{A.2}$$

Here and throughout Appendix A, for simplicity of notation, we use the notation h_x^0 for $h_x(x,y)$ and h_x for $h_x(u,v)$, and the same applies for other partial derivatives of h.

Differentiating these two equations with respect to z, we can solve for $u_z(z)$ and $v_z(z)$:

$$u_z(z)=\frac{h_x^0}{h_x^0h_x+h_y^0h_y}, \tag{A.3}$$

$$v_z(z)=\frac{h_y^0}{h_x^0h_x+h_y^0h_y}. \tag{A.4}$$

Differentiating these derivatives with respect to z, we have

$$u_{zz}(z)=-\frac{(h_x^0)^2[h_{xx}u_z(z)+h_{xy}v_z(z)]+h_x^0h_y^0[h_{xy}u_z(z)+h_{yy}v_z(z)]}{(h_x^0h_x+h_y^0h_y)^2}, \tag{A.5}$$

$$v_{zz}(z)=-\frac{h_x^0h_y^0[h_{xx}u_z(z)+h_{xy}v_z(z)]+(h_y^0)^2[h_{xy}u_z(z)+h_{yy}v_z(z)]}{(h_x^0h_x+h_y^0h_y)^2}. \tag{A.6}$$

Since $(u,v)\to(x,y)$ as $z\to h_n$, we have

$$u_z(h_n)=\frac{h_x^0}{(h_x^0)^2+(h_y^0)^2}, \tag{A.7}$$

$$v_z(h_n)=\frac{h_y^0}{(h_x^0)^2+(h_y^0)^2}, \tag{A.8}$$

and

$$u_{zz}(h_n)=-\frac{h_x^0[(h_x^0)^2h_{xx}^0+2h_x^0h_y^0h_{xy}^0+(h_y^0)^2h_{yy}^0]}{[(h_x^0)^2+(h_y^0)^2]^3}, \tag{A.9}$$

$$v_{zz}(h_n)=-\frac{h_y^0[(h_x^0)^2h_{xx}^0+2h_x^0h_y^0h_{xy}^0+(h_y^0)^2h_{yy}^0]}{[(h_x^0)^2+(h_y^0)^2]^3}. \tag{A.10}$$

The signed distance from the point $\mathbf{P}$ to the point $\mathbf{X}$, which is positive (or negative) if $\mathbf{XP}$ is in the positive (or negative) direction of $-\nabla h(x,y)$, is given by

$$d=-\frac{(u-x)h_x^0+(v-y)h_y^0}{\sqrt{(h_x^0)^2+(h_y^0)^2}}. \tag{A.11}$$

Using (A.7)–(A.10) and Taylor expansions of u at x and v at y up to second order of $z - h_n$, we have

$$d = -\frac{z-h_n}{[(h_x^0)^2+(h_y^0)^2]^{1/2}} + \frac{(h_x^0)^2 h_{xx}^0 + 2h_x^0 h_y^0 h_{xy}^0 + (h_y^0)^2 h_{yy}^0}{2[(h_x^0)^2+(h_y^0)^2]^{5/2}}(z-h_n)^2 + O((z-h_n)^3) \tag{A.12}$$

as $z \to h_n$.

Appendix B. Proof of Theorem 4.1. In this appendix, we give the proof of Theorem 4.1 on the error estimate of the trapezoidal rule for the integral of some singular function. We will use the following theorems, whose proofs can be found in Sidi and Israeli [40] or the references therein.

In the theorems below, interval $[a,b]$ is divided into m subintervals with $\Delta x = (b-a)/m$, $x_j = a + (j-1)\Delta x$, $j = 1, \ldots, m+1$. It is assumed that $g(x)$ is $2N$ times continuously differentiable on $[a,b]$ throughout Appendix B.

THEOREM B.1 (Euler–Maclaurin formula).

$$\begin{aligned}\int_a^b g(x)\,dx = \Delta x \left(\frac{g(a)+g(b)}{2} + \sum_{j=2}^{m} g(x_j) \right)\\ + \sum_{k=1}^{N-1} \frac{B_{2k}}{(2k)!}[g^{(2k-1)}(a) - g^{(2k-1)}(b)]\Delta x^{2k} + O(\Delta x^{2N}),\end{aligned} \tag{B.1}$$

where B_k's are the Bernoulli numbers.

THEOREM B.2 (Sidi and Israeli [40]). *Let $G(x) = g(x)/(x-t)$ with $t = x_{j_0}$ for some j_0. Then*

$$\begin{aligned}\int_a^b G(x)\,dx = \Delta x \left(\frac{G(a)+G(b)}{2} + \sum_{2\le j\le m, j\ne j_0} G(x_j) \right) + \Delta x g'(t)\\ + \sum_{k=1}^{N-1} \frac{B_{2k}}{(2k)!}[G^{(2k-1)}(a) - G^{(2k-1)}(b)]\Delta x^{2k} + O(\Delta x^{2N}),\end{aligned} \tag{B.2}$$

where B_k's are the Bernoulli numbers.

THEOREM B.3 (Sidi and Israeli [40]). *Let $G(x) = g(x)\log|x-t|$ with $t = x_{j_0}$ for some j_0. Then*

$$\begin{aligned}\int_a^b G(x)\,dx = \Delta x \left(\frac{G(a)+G(b)}{2} + \sum_{2\le j\le m, j\ne j_0} G(x_j) \right) + g(t)\log(\Delta x)\Delta x\\ + \sum_{k=1}^{N-1} \frac{B_{2k}}{(2k)!}[G^{(2k-1)}(a) - G^{(2k-1)}(b)]\Delta x^{2k}\\ +2\sum_{k=0}^{N-1} \frac{\zeta'(-2k)}{(2k)!} g^{(2k)}(t)\Delta x^{2k+1} + O(\Delta x^{2N}),\end{aligned} \tag{B.3}$$

where B_k's are the Bernoulli numbers and $\zeta(\tau)$ is the Riemann zeta function.

Proof of Theorem 4.1. Using the above theorems, we have

$$\int_a^b G(x)\,dx - \Delta x\left(\frac{G(a)+G(b)}{2} + \sum_{2\le j\le m, j\ne j_0} G(x_j)\right)$$

$$= g_1(t)\Delta x\log(\Delta x) - \log(2\pi)g_1(t)\Delta x + g_2'(t)\Delta x + g_3(t)\Delta x + O(\Delta x^2). \tag{B.4}$$

On the other hand,

$$\begin{aligned}
&\int_{t-\frac{\Delta x}{2}}^{t+\frac{\Delta x}{2}} G(x)\,dx\\
&= \int_{t-\frac{\Delta x}{2}}^{t+\frac{\Delta x}{2}} \left[g_1(x)\log|x-t| + \frac{g_2(x)}{x-t} + g_3(x)\right] dx\\
&= \int_{t-\frac{\Delta x}{2}}^{t+\frac{\Delta x}{2}} \Bigg[(g_1(t)+g_1'(t)(x-t)+O((x-t)^2))\log|x-t|\\
&\qquad + \frac{g_2(t)+g_2'(t)(x-t)+O((x-t)^2)}{x-t} + g_3(t) + O(x-t)\Bigg] dx\\
&= \int_{t-\frac{\Delta x}{2}}^{t+\frac{\Delta x}{2}} [g_1(t)\log|x-t| + g_2'(t) + g_3(t)]\,dx + O(\Delta x^2)
\end{aligned}$$

$$= g_1(t)\Delta x\log(\Delta x) - (1+\log 2)g_1(t)\Delta x + g_2'(t)\Delta x + g_3(t)\Delta x + O(\Delta x^2). \tag{B.5}$$

A combination of these two equations gives the theorem.

REFERENCES

[1] O. L. Alerhand, D. Vanderbilt, R. D. Meade, and J. D. Joannopoulos, *Spontaneous formation of stress domains on crystal surfaces*, Phys. Rev. Lett., 61 (1988), pp. 1973–1976.

[2] R. J. Asaro and W. A. Tiller, *Interface morphology development during stress-corrosion cracking: Part* I. *Via surface diffusion*, Metall. Trans., 3 (1972), pp. 1789–1796.

[3] W. K. Burton, N. Cabrera, and F. Frank, *The growth of crystals and the equilibrium structures of their surfaces*, Phil. Trans. Roy. Soc., 243 (1951), pp. 299–358.

[4] C. H. Chiu and H. Gao, *Stressed singularities along a cycloid rough surface*, Internat. J. Solids Structures, 30 (1993), pp. 2983–3012.

[5] C. Duport, P. Nozieres, and J. Villain, *New instability in molecular-beam epitaxy*, Phys. Rev. Lett., 74 (1995), pp. 134–137.

[6] C. Duport, P. Politi, and J. Villain, *Growth instabilities induced by elasticity in a vicinal surface*, J. Phys. I, 5 (1995), pp. 1317–1350.

[7] W. E, *Dynamics of vortices in Ginzburg–Landau theories with applications to superconductivity*, Phys. D, 77 (1994), pp. 383–404.

[8] L. B. Freund and F. Jonsdottir, *Instability of a biaxially stressed thin film on a substrate due to material diffusion over its free surface*, J. Mech. Phys. Solids, 41 (1993), pp. 1245–1264.

[9] L. B. Freund and S. Suresh, *Thin Film Materials: Stress, Defect Formation, and Surface Evolution*, Cambridge University Press, Cambridge, UK, 2003.

[10] H. Gao and W. D. Nix, *Surface roughening of heteroepitaxial thin films*, Annu. Rev. Mater. Sci., 29 (1999), pp. 173–209.

[11] M. A. Grinfeld, *Instability of the separation boundary between a non-hydrostatically stressed elastic body and a melt*, Soviet Phys. Dokl., 31 (1986), pp. 831–834.

[12] F. Haußer, M. E. Jabbour, and A. Voigt, *A step-flow for the heteroepitaxial growth of strained, substitutional, binary alloy films with phase segregation:* I. *Theory*, Multiscale Model. Simul., 6 (2007), pp. 158–189.

[13] C. Herring, *The use of classical macroscopic concepts in surface energy problems*, in Structure and Properties of Solid Surfaces, R. Gomer and C. S. Smith, eds., University of Chicago Press, Chicago, 1953, pp. 5–72.

[14] J. P. Hirth and J. Lothe, *Theory of Dislocations*, 2nd ed., John Wiley & Sons, New York, 1982.

[15] B. Houchmandzadeh and C. Misbah, *Elastic interaction between modulated steps on a vicinal surface*, J. Phys. I, 5 (1995), pp. 685–698.

[16] N. Israeli and D. Kandel, *Profile of a decaying crystalline cone*, Phys. Rev. B, 60 (1999), pp. 5946–5962.

[17] K. L. Johnson, *Contact Mechanics*, Cambridge University Press, Cambridge, UK, 1985.

[18] V. M. Kaganer and K. H. Ploog, *Energies of strained vicinal surfaces and strained islands*, Phys. Rev. B, 64 (2001), article 205301.

[19] R. V. Kukta and K. Bhattacharya, *A 3-D model of step flow mediated crystal growth under the combined influences of stress and diffusion*, Thin Solid Films, 357 (1999), pp. 35–39.

[20] R. V. Kukta and K. Bhattacharya, *A micromechanical model of surface steps*, J. Mech. Phys. Solids, 50 (2002), pp. 615–649.

[21] F. Lançon and J. Villain, *Dynamics of crystal surface below its roughening transition*, Phys. Rev. Lett., 64 (1990), pp. 293–296.

[22] F. Léonard and J. Tersoff, *Competing step instabilities at surfaces under stress*, Appl. Phys. Lett., 83 (2003), pp. 72–74.

[23] F. Liu, J. Tersoff, and M. G. Lagally, *Self-organization of steps in growth of strained films on vicinal substrates*, Phys. Rev. Lett., 80 (1998), pp. 1268–1271.

[24] V. I. Marchenko and A. Ya. Parshin, *Elastic properties of the crystal surface*, Soviet Phys. JETP, 52 (1980), pp. 129–131.

[25] D. Margetis and R. V. Kohn, *Continuum relaxation of interacting steps on crystal surfaces in $2+1$ dimensions*, Multiscale Model. Simul., 5 (2006), pp. 729–758.

[26] D. Margetis, M. J. Aziz, and H. A. Stone, *Continuum description of profile scaling in nanostructure decay*, Phys. Rev. B, 69 (2004), article 041404(R).

[27] D. Margetis, M. J. Aziz, and H. A. Stone, *Continuum approach to self-similarity and scaling in morphological relaxation of a crystal with a facet*, Phys. Rev. B, 71 (2005), article 165432.

[28] D. Margetis, P. W. Fok, M. J. Aziz, and H. A. Stone, *Continuum theory of nanostructure decay via a microscale condition*, Phys. Rev. Lett., 97 (2006), article 096102.

[29] P. Müller and A. Saúl, *Elastic effects on surface physics*, Surf. Sci. Rep., 54 (2004), pp. 157–258.

[30] W. W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.

[31] M. Ozdemir and A. Zangwill, *Morphological equilibration of a corrugated crystalline surface*, Phys. Rev. B, 42 (1990), pp. 5013–5024.

[32] S. Paulin, F. Gillet, O. Pierre-Louis, and C. Misbah, *Unstable step meandering with elastic interactions*, Phys. Rev. Lett., 86 (2001), pp. 5538–5541.

[33] A. Pimpinelli and J. Villain, *Physics of Crystal Growth*, Cambridge University Press, Cambridge, UK, 1998.

[34] L. M. Pismen and J. Rubinstein, *Motion of vortex lines in the Ginzburg–Landau model*, Phys. D, 47 (1991), pp. 353–360.

[35] W. Selke and P. M. Duxbury, *Equilibration of crystal surfaces*, Phys. Rev. B, 52 (1995), pp. 17468–17479.

[36] A. Ramasubramaniam and V. B. Shenoy, *Three-dimensional simulations of self-assembly of hut-shaped Si-Ge quantum dots*, J. Appl. Phys., 95 (2004), pp. 7813–7824.

[37] J. M. Rickman and D. J. Srolovitz, *Defect interactions on solid surfaces*, Surf. Sci., 284 (1993), pp. 211–221.

[38] V. B. Shenoy, *Growth of epitaxial nanowires by controlled coarsening of strained islands*, Appl. Phys. Lett., 85 (2004), pp. 2376–2378.

[39] V. B. Shenoy and L. B. Freund, *A continuum description of the energetics and evolution of stepped surfaces in strained nanostructures*, J. Mech. Phys. Solids, 50 (2002), pp. 1817–1841.

[40] A. Sidi and M. Israeli, *Quadrature methods for periodic singular and weakly singular Fredholm integral equations*, J. Sci. Comput., 3 (1988), pp. 201–231.

[41] B. J. Spencer and D. I. Meiron, *Nonlinear evolution of the stress-driven morphological instability in a 2-dimensional semi-infinite solid*, Acta Metall. Mater., 42 (1994), pp. 3629–3641.

[42] B. J. Spencer, P. W. Voorhees, and S. H. Davis, *Morphological instability in epitaxially strained dislocation-free solid films*, Phys. Rev. Lett., 67 (1991), pp. 3696–3699.

[43] D. J. Srolovitz, *On the stability of surfaces of stressed solids*, Acta Metall., 37 (1989), pp. 621–625.

[44] J. Stewart, O. Pohland, and J. M. Gibson, *Elastic-displacement field of an isolated surface step*, Phys. Rev. B, 49 (1994), pp. 13848–13858.

[45] Z. Suo, *Motion of microscopic surfaces in materials*, Adv. Appl. Mech., 33 (1997), pp. 193–294.

[46] J. Tersoff and E. Pehlke, *Sinuous step instability on the Si*(001) *surface*, Phys. Rev. Lett., 68 (1992), pp. 816–819.

[47] J. Tersoff, Y. H. Phang, Z. Zhang, and M. G. Lagally, *Step-bunching instability of vicinal surfaces under stress*, Phys. Rev. Lett., 75 (1995), pp. 2730–2733.

[48] R. M. Tromp and M. C. Reuter, *Wavy steps on Si*(001), Phys. Rev. Lett., 68 (1992), pp. 820–822.

[49] Y. Xiang, *Derivation of a continuum model for epitaxial growth with elasticity on vicinal surface*, SIAM J. Appl. Math., 63 (2002), pp. 241–258.

[50] Y. Xiang and W. E, *Nonlinear evolution equation for the stress-driven morphological instability*, J. Appl. Phys., 91 (2002), pp. 9414–9422.

[51] Y. Xiang and W. E, *Misfit elastic energy and a continuum model for epitaxial growth with elasticity*, Phys. Rev. B, 69 (2004), article 035409.

[52] W. H. Yang and D. J. Srolovitz, *Crack-like surface instabilities in stressed solids*, Phys. Rev. Lett., 71 (1993), pp. 1593–1596.

[53] W. H. Yang and D. J. Srolovitz, *Surface-morphology evolution in stressed solids: Surface diffusion controlled crack initiation*, J. Mech. Phys. Solids, 42 (1994), pp. 1551–1574.

[54] Y. W. Zhang and A. F. Bower, *Numerical simulations of island formation in a coherent strained epitaxial thin film system*, J. Mech. Phys. Solids, 47 (1999), pp. 2273–2297.

[55] X. Zhu, H. Xu, and Y. Xiang, *A continuum model for the long-range elastic interaction on stepped epitaxial surfaces in* $2+1$ *dimensions*, Phys. Rev. B, submitted.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1415–1437

VOLTAGE AND CURRENT SPECTRA FOR MATRIX POWER CONVERTERS*

STEPHEN M. COX† AND STEPHEN C. CREAGH†

Abstract. Matrix power converters are used for transforming one alternating-current power supply to another, with different peak voltage and frequency. There are three input lines, with sinusoidally varying voltages which are 120° out of phase one from another, and the output is to be delivered as a similar three-phase supply. The matrix converter switches rapidly, to connect each output line in sequence to each of the input lines in an attempt to synthesize the prescribed output voltages. The switching is carried out at high frequency and it is of practical importance to know the frequency spectra of the output voltages and of the input and output currents. We determine in this paper these spectra using a new method, which has significant advantages over the prior default method (a multiple Fourier series technique), leading to a considerably more direct calculation. In particular, the determination of the input current spectrum is feasible here, whereas it would be a significantly more daunting procedure using the prior method instead.

Key words. matrix power converter, power electronics, Fourier spectrum

AMS subject classifications. 42A16, 94C05

DOI. 10.1137/080718863

1. Introduction.

In electrical and electronic engineering, there are many applications in which it is necessary to convert a power supply from one voltage and frequency to another. Particular examples arise in aeronautical and marine applications, since there are increasingly many electrically powered devices aboard aircraft and ships, all with separate demands in terms of power supplies. The field of power conversion, while of great economic importance, thus poses particular technological challenges in aircraft in particular, where it is clearly highly desirable that power conversion be achieved without recourse to heavy bulk energy storage elements.

In modern solid-state power converters, the need for intermediate energy storage is avoided, because the output voltage is generated by rapidly switching between multiple input voltages (see, for example, [7]). The aim is that the low-frequency components of the output synthesize a prescribed waveform, while the high-frequency components related to the switching are ultimately filtered out.

In this paper, we describe a compact means of determining the voltage and current spectra for one such application of particular technological significance: the matrix power converter [12]. This is a device which aims to convert an alternating-current power supply at one voltage and frequency to a second at a different voltage and frequency. Applications of matrix converters include adjustable-speed drives, where the speed of the motor is governed by the frequency of its power supply. A significant benefit of the approach outlined in this paper is that we are able to give explicit and detailed descriptions of input currents, which are considerably more complex and difficult to determine than the output voltages and currents which have been predominantly studied in the past.

*Received by the editors March 19, 2008; accepted for publication (in revised form) December 3, 2008; published electronically February 19, 2009. This work was supported by the Engineering and Physical Sciences Research Council (grant EP/E018580/1).
http://www.siam.org/journals/siap/69-5/71886.html

†School of Mathematical Sciences, University of Nottingham, Nottingham NG7 2RD, UK (stephen.cox@nottingham.ac.uk, stephen.creagh@nottingham.ac.uk).

The matrix converter switching frequency greatly exceeds the input and output frequencies. The way in which its switching takes place is termed its modulation strategy here, and there are many such strategies adopted in practice, of which a comparatively simple variant is comprehensively analyzed in this paper. The matrix converter successively connects, via switches, each output line to each of the input lines, according to the modulation strategy. Thus the voltage on any given output line comprises short segments of the three input sinusoidal waveforms; it contains both low-frequency contributions (from the input voltages) and high-frequency contributions (from the switching). It is the spectrum of the output voltages and currents that we compute here, along with the more involved calculation for the corresponding currents drawn from the input lines.

In digital implementations, the input voltages are measured (sampled) at high frequency, at the start of each switching period. Then after each sample a calculation must be done to determine the corresponding switching times to achieve the desired output. This leads to so-called *regular* or *uniform* sampling of the input to determine the modulation strategy. The delay between sampling and switching results in undesirable distortion in the form of unintended low-frequency components in the output [5]; it also affects the high-frequency part of the spectrum, but this is not so serious provided that the low-pass filtering still effectively removes such components.

Although less relevant to the power converter application, an alternative sampling technique is also analyzed here: so-called *natural* sampling, which is widely used in, for example, audio applications [3, 6, 8]. In natural sampling, an analogue device compares one of the input voltages with some reference waveform and switches whenever the two become (instantaneously) equal. The lack of delays in natural sampling leads to a more accurate spectrum for the audio component of the output [7]; a comparison between the spectra for regular and natural sampling allows us to determine what aspects of the former spectrum are due to associated digital implementation effects.

In the engineering literature, spectra for switching devices are generally computed by a multiple Fourier series method usually ascribed to Black [3], but acknowledged to go at least as far back as Bennett [2] (see, for example, [4]). The method involves introducing separate independent variables representing time scaled by each of the input, output, and switching frequencies, then writing the required quantities as multiple Fourier series, in terms of each of these variables separately. The corresponding Fourier coefficients are then computed. Finally, the answer is specialized to the physical case, in which the separate time variables are recognized to be constant multiples of one another. The method is simplest for natural sampling, but can be modified for regular sampling, although it is more algebraically involved in that case.

The major content of this paper is the development of more direct methods than Black's for determining the output voltage, output current, and input current spectra. The methods contained herein can be used for regular and natural sampling, although the order in which various steps are applied is different in the two cases if the greatest efficiency is to be achieved. However, in contrast to Black's method, neither calculation is intrinsically more algebraically cumbersome than the other. Furthermore, our analysis, although presented here for the matrix converter problem, is in fact readily adaptable to any other switching problem for which Black's method is the usual default, for example the modeling of class-D audio amplifiers [6].

In section 2 we outline some notation and describe model calculations (given for both regular and natural sampling), which form the building blocks for many of the subsequent calculations. The output voltages are then computed, using these building-block solutions, for both types of sampling. In section 3 we introduce further notation and calculate the output and input currents, for general output impedances, illustrating our results in section 4. In section 5 we show how to derive more rapidly convergent solutions when the form of the output impedances is known (it is often the case that the output loads may be approximated by a resistor and an inductor in series, for example), considerably reducing the computation time. Our conclusions are given in section 6. In the appendix, we illustrate how the results in this paper may be extended to a more complicated modulation strategy.

2. Calculation of output voltages. We begin by calculating the spectra of output voltages in a matrix converter. Some of the main ideas of this simpler calculation recur in the more involved calculation for input currents and it is useful to set out the main features and establish notation in the simpler context.

2.1. Notation for voltages. In an idealized matrix converter [12] (see Figure 2.1), there are three input voltages, which we label

$$v^A(t) = \mathrm{e}^{\mathrm{i}\omega_0 t}, \qquad v^B(t) = \mathrm{e}^{\mathrm{i}(\omega_0 t + 2\pi/3)}, \qquad v^C(t) = \mathrm{e}^{\mathrm{i}(\omega_0 t + 4\pi/3)}, \tag{2.1}$$

where ω_0 is the input frequency and voltages are scaled to give unit peak input voltages. Of course, the physical voltages are the real parts of the expressions given in (2.1). It will be convenient to write these collectively as a vector

$$\mathbf{v}^{\mathrm{in}}(t) = \begin{pmatrix} v^A(t) \\ v^B(t) \\ v^C(t) \end{pmatrix} = \mathrm{e}^{\mathrm{i}\omega_0 t} \begin{pmatrix} 1 \\ p \\ p^2 \end{pmatrix}, \qquad \text{where} \quad p = \mathrm{e}^{2\pi\mathrm{i}/3}.$$

Three output voltages, denoted $v^a(t)$, $v^b(t)$, and $v^c(t)$ and written in the vector form,

$$\mathbf{v}^{\mathrm{out}}(t) = \begin{pmatrix} v^a(t) \\ v^b(t) \\ v^c(t) \end{pmatrix},$$

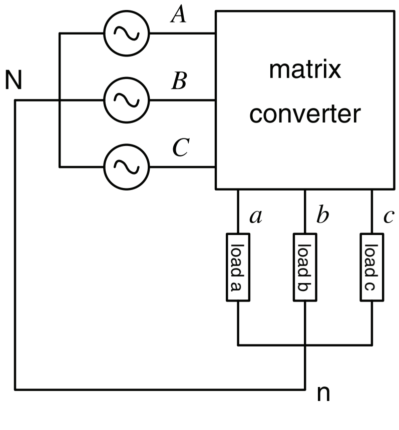


FIG. 2.1. *Diagram of a matrix converter. Three input lines (A, B, and C) each supply sinusoidal voltages, each* 120° *out of phase with any other. There are loads on each of the output lines (a, b, and c). The neutral point of the input lines is denoted by* N *and that of the output lines by* n*; these neutral points are assumed to be connected to one another, and to be at a nominal zero volts.*

are obtained by sampling the input voltages during intervals which repeat quasiperiodically, according to the modulation strategy. We scale time so that the switching period is unity (hence the switching frequency is 2π), and consider a simple form of switching pattern such that each output, labeled $o = a, b, c$, is of the form [9]

$$v^o(t) = \begin{cases} v^A(t) & \text{for } n < t < n + \alpha_n^o, \\ v^B(t) & \text{for } n + \alpha_n^o < t < n + \beta_n^o, \\ v^C(t) & \text{for } n + \beta_n^o < t < n + 1 \end{cases} \tag{2.2}$$

(more sophisticated switching strategies can also be analyzed using the methods described in this paper; these are discussed briefly in the appendix). An important assumption in our analysis will be that the switching frequency is much larger than the input frequency: $2\pi \gg \omega_0$. This is certainly the case in practical implementations, to allow the high-frequency switching components to be filtered without affecting the desired low-frequency components.

We encode the relationships in (2.2) using a switching matrix

$$M(t) = \begin{pmatrix} F^{aA}(t) & F^{aB}(t) & F^{aC}(t) \\ F^{bA}(t) & F^{bB}(t) & F^{bC}(t) \\ F^{cA}(t) & F^{cB}(t) & F^{cC}(t) \end{pmatrix},$$

whose elements are either 0 or 1 at any given instant, according to which input and output lines are connected. Then

$$\mathbf{v}^{\text{out}}(t) = M(t)\mathbf{v}^{\text{in}}(t). \tag{2.3}$$

For example,

$$F^{aA}(t) = \sum_{n=-\infty}^{\infty} \psi_{n,n+\alpha_n^a}(t), \tag{2.4}$$

where the step function ψ_{t_1,t_2} is defined by

$$\psi_{t_1,t_2}(t) = \begin{cases} 1 & \text{if } t_1 < t < t_2, \\ 0 & \text{otherwise,} \end{cases}$$

and similar expressions can be written for the other elements of $M(t)$, using (2.2).

2.2. A model calculation for the case of uniform sampling. Before describing the full calculation of the three-phase output voltages it is useful to outline a model calculation which illustrates the essence of our approach in a somewhat simpler setting. We consider a function,

$$F(t) = \sum_{n=-\infty}^{\infty} \psi_{n+\alpha_n,n+\beta_n}(t), \tag{2.5}$$

which samples a unit input voltage over the quasiperiodically repeating intervals

$$n + \alpha_n < t < n + \beta_n.$$

In the case of *uniform sampling*, we assume that the switching times are determined by sampling the continuous functions

$$\alpha(\tau) = \mu_\alpha + \lambda_\alpha \cos(\Omega\tau + \delta_0) \quad \text{and} \quad \beta(\tau) = \mu_\beta + \lambda_\beta \cos(\Omega\tau + \delta_0 + \delta_1) \tag{2.6}$$

at the evenly spaced times $\tau = n$ [1, 9, 10]. In other words,

$$\alpha_n = \alpha(n) \qquad \text{and} \qquad \beta_n = \beta(n). \tag{2.7}$$

Note that the matrix product in (2.3) consists of sums of functions of the form (2.5), modulated by the harmonic function $\mathrm{e}^{\mathrm{i}\omega_0 t}$.

Adopting the convention that the Fourier transform is written as

$$\hat{F}(\omega) = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega t} F(t)\,\mathrm{d}t,$$

and noting that the Fourier transform of the step function $\psi_{t_1,t_2}(t)$ is

$$\hat{\psi}_{t_1,t_2}(\omega) = \frac{\mathrm{e}^{-\mathrm{i}\omega t_2} - \mathrm{e}^{-\mathrm{i}\omega t_1}}{-\mathrm{i}\omega},$$

we find

$$\hat{F}(\omega) = \sum_{n=-\infty}^{\infty} \mathrm{e}^{-in\omega} \hat{\psi}_{\alpha_n,\beta_n}(\omega) = \sum_{n=-\infty}^{\infty} \mathrm{e}^{-in\omega} \frac{\mathrm{e}^{-\mathrm{i}\omega\beta_n} - \mathrm{e}^{-\mathrm{i}\omega\alpha_n}}{-\mathrm{i}\omega}.$$

We now make use of the identity [11]

$$\mathrm{e}^{-\mathrm{i}z\cos\theta} = \sum_{m=-\infty}^{\infty} \mathrm{J}_m(z)(-\mathrm{i})^m \mathrm{e}^{\mathrm{i}m\theta}, \tag{2.8}$$

so that

$$\frac{\mathrm{e}^{-\mathrm{i}\omega\beta_n} - \mathrm{e}^{-\mathrm{i}\omega\alpha_n}}{-\mathrm{i}\omega} = \sum_{m=-\infty}^{\infty} X_m(\omega) \mathrm{e}^{\mathrm{i}nm\Omega + \mathrm{i}m\delta_0},$$

where, using (2.6), we find

$$X_m(\omega) = \frac{(-\mathrm{i})^m}{-\mathrm{i}\omega} \left(\mathrm{e}^{-\mathrm{i}\omega\mu_\beta + \mathrm{i}m\delta_1} \mathrm{J}_m(\omega\lambda_\beta) - \mathrm{e}^{-\mathrm{i}\omega\mu_\alpha} \mathrm{J}_m(\omega\lambda_\alpha) \right). \tag{2.9}$$

It is useful to record the following limiting values:

$$X_m(0) = \begin{cases} \mu_\beta - \mu_\alpha & \text{when } m = 0, \\ \frac{1}{2}\left(\mathrm{e}^{m\mathrm{i}\delta_1}\lambda_\beta - \lambda_\alpha\right) & \text{when } m = \pm 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Then, using the Poisson summation formula in the form

$$\sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}n(\omega - m\Omega)} = 2\pi \sum_{n=-\infty}^{\infty} \delta(\omega - m\Omega - 2\pi n),$$

we find that

$$\begin{aligned} \hat{F}(\omega) &= \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m(\omega)\, \mathrm{e}^{-\mathrm{i}n(\omega - m\Omega) + \mathrm{i}m\delta_0} \\ &= 2\pi \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m(\omega_{nm}) \mathrm{e}^{m\mathrm{i}\delta_0} \delta(\omega - \omega_{nm}), \end{aligned} \tag{2.11}$$

where we denote

$$\omega_{nm} = 2\pi n + m\Omega. \tag{2.12}$$

The result (2.11) can alternatively be written in the time domain as

$$F(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m(\omega_{nm}) \mathrm{e}^{\mathrm{i}\omega_{nm}t + \mathrm{i}m\delta_0}. \tag{2.13}$$

We therefore find that the spectrum of the function $F(t)$ is confined to integer combinations of the switching frequency 2π and the modulation frequency Ω (cf. [3]). Furthermore, the frequencies of particular practical interest are those with $n = 0$ (those with $n \neq 0$ will be filtered out), and it is easily verified that in the limit $\Omega \ll 2\pi$ the dominant frequencies with $n = 0$ are $\omega_{00} = 0$ and $\omega_{0\pm1} = \pm\Omega$. It will later prove useful to denote by $F_0(t)$ the terms in (2.13) for $n = 0$; summing over m the contributions to $F(t)$ with $n = 0$ is then easily seen to give

$$F_0(t) \equiv \sum_{m=-\infty}^{\infty} X_m(\omega_{0m}) \mathrm{e}^{\mathrm{i}\omega_{0m}t + \mathrm{i}m\delta_0} = \beta(t) - \alpha(t) + O(\Omega). \tag{2.14}$$

By comparing (2.14) with (2.6), we see that the $n = 0$ contribution $F_0(t)$ is thus, with errors of order Ω, a sinusoidal signal with frequency Ω, plus a constant signal.

2.3. Output voltages in the case of uniform sampling. The model calculation in section 2.2 can now be used as the basis for a more complete description of the output voltages. We begin by describing more explicitly the switching conventions in (2.2). These are designed to generate output voltages

$$\mathbf{v}_{\mathrm{ref}}^{\mathrm{out}}(t) = \begin{pmatrix} v_{\mathrm{ref}}^a(t) \\ v_{\mathrm{ref}}^b(t) \\ v_{\mathrm{ref}}^c(t) \end{pmatrix} = q\mathrm{e}^{\mathrm{i}\omega_1 t} \begin{pmatrix} 1 \\ p \\ p^2 \end{pmatrix}, \tag{2.15}$$

where ω_1 is the output frequency and q is the output amplitude. The subscript "ref" indicates that the corresponding quantity is the intended, reference state; the actual output voltage will generally approximate this reference value in its low-frequency spectrum, but also contain slight low-frequency distortion terms and significant high-frequency components. The matching of the low-frequency terms in $\mathbf{v}^{\mathrm{out}}(t)$ and $\mathbf{v}_{\mathrm{ref}}^{\mathrm{out}}(t)$ is achieved by letting α_n^o and β_n^o in (2.2) oscillate with an appropriate frequency that is much smaller than the switching frequency.

We focus on the simplest Venturini switching [1, 9, 10], in which, for each output $o = a$, b, or c, the times α_n^o and β_n^o are obtained by sampling smooth functions as in (2.6) and (2.7), with

$$\alpha^o(\tau) = \tfrac{1}{3} + \tfrac{2}{3}q\cos(\Omega\tau + \delta^o), \qquad \beta^o(\tau) = \tfrac{2}{3} + \tfrac{2}{3}q\cos(\Omega\tau + \delta^o - \tfrac{\pi}{3}). \tag{2.16}$$

Here,

$$\Omega \equiv \omega_1 - \omega_0$$

is the difference between output and input frequencies and

$$\delta^a = 0, \quad \delta^b = \tfrac{2}{3}\pi, \quad \text{and} \quad \delta^c = \tfrac{4}{3}\pi.$$

The coefficients in (2.16) were originally derived by attempting to generate the correct low-frequency components in the output voltages, in the limit $\Omega \to 0$.

We now adapt the model calculation in section 2.2 to describe the output voltages. Note that (2.3) indicates that any given output voltage can be written as

$$\begin{aligned} v^o(t) &= \mathrm{e}^{\mathrm{i}\omega_0 t} \sum_{n=-\infty}^{\infty} \psi_{n,n+\alpha_n^o}(t) + p\psi_{n+\alpha_n^o,n+\beta_n^o}(t) + p^2\psi_{n+\beta_n^o,n+1}(t) \\ &= \left(F^{oA}(t) + pF^{oB}(t) + p^2F^{oC}(t)\right)\mathrm{e}^{\mathrm{i}\omega_0 t}, \qquad o = a, b, c, \end{aligned}$$

where

$$\begin{aligned} F^{oA}(t) &= \sum_{n=-\infty}^{\infty} \psi_{n,n+\alpha_n^o}(t), \\ F^{oB}(t) &= \sum_{n=-\infty}^{\infty} \psi_{n+\alpha_n^o,n+\beta_n^o}(t), \qquad (2.17) \\ F^{oC}(t) &= \sum_{n=-\infty}^{\infty} \psi_{n+\beta_n^o,n+1}(t) \end{aligned}$$

are all particular cases of the function $F(t)$ described in section 2.2. Repeating the calculations there, we find that $F^{oA}(t)$, $F^{oB}(t)$, and $F^{oC}(t)$ have Fourier transforms

$$(2.18) \qquad \hat{F}^{oi}(\omega) = 2\pi \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m^i(\omega_{nm})\mathrm{e}^{m\mathrm{i}\delta^o}\delta(\omega - \omega_{nm}), \qquad i = A, B, C,$$

where

$$\begin{aligned} X_m^A(\omega) &= \frac{(-\mathrm{i})^m}{-\mathrm{i}\omega}\left[\mathrm{e}^{-\mathrm{i}\omega/3}\mathrm{J}_m\left(\tfrac{2}{3}q\omega\right) - \delta_{m0}\right], \\ X_m^B(\omega) &= \frac{(-\mathrm{i})^m}{-\mathrm{i}\omega}\left[\mathrm{e}^{-2\mathrm{i}\omega/3-\mathrm{i}m\pi/3} - \mathrm{e}^{-\mathrm{i}\omega/3}\right]\mathrm{J}_m\left(\tfrac{2}{3}q\omega\right), \qquad (2.19) \\ X_m^C(\omega) &= \frac{(-\mathrm{i})^m}{-\mathrm{i}\omega}\left[\mathrm{e}^{-\mathrm{i}\omega}\delta_{m0} - \mathrm{e}^{-2\mathrm{i}\omega/3-\mathrm{i}m\pi/3}\mathrm{J}_m\left(\tfrac{2}{3}q\omega\right)\right]; \end{aligned}$$

here δ_{0m} is the Kronecker δ. Note that the quantities $X_m^i(\omega)$ do not depend on the output line.

The transformed output voltages are given by

$$\hat{v}^o(\omega) = \hat{F}^{oA}(\omega - \omega_0) + p\hat{F}^{oB}(\omega - \omega_0) + p^2\hat{F}^{oC}(\omega - \omega_0),$$

which can then be written as

$$\hat{v}^o(\omega) = 2\pi \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}m\delta^o} V_{nm}\delta(\omega - \Omega_{nm}),$$

where we denote

$$V_{nm} = X_m^A(\omega_{nm}) + pX_m^B(\omega_{nm}) + p^2X_m^C(\omega_{nm})$$

and

$$\Omega_{nm} = \omega_0 + \omega_{nm} = \omega_0 + 2\pi n + m\Omega.$$

The corresponding expression in the time domain gives us the key result of this section, that the output voltages may be written as

$$v^o(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} V_{nm} \mathrm{e}^{\mathrm{i}\Omega_{nm} t + \mathrm{i} m \delta^o}. \tag{2.20}$$

Note that the amplitudes V_{nm} are common to all three output lines; the differences between the three output voltages result from the different values of δ^o for $o = a, b, c$ in (2.20).

The most physically interesting part of the result is

$$v_0^o(t) \equiv \sum_{m=-\infty}^{\infty} V_{0m} \mathrm{e}^{i\Omega_{0m} t + \mathrm{i} m \delta^o} = \sum_{m=-1}^{1} V_{0m} \mathrm{e}^{i\Omega_{0m} t + \mathrm{i} m \delta^o} + O(\Omega). \tag{2.21}$$

In analogy with (2.10), we find the limiting cases

$$X_0^i(0) = \tfrac{1}{3}, \qquad i = A, B, C$$

and

$$X_{\pm 1}^A(0) = \tfrac{1}{3} q, \qquad X_{\pm 1}^B(0) = \tfrac{1}{3} q (\mathrm{e}^{\mp \mathrm{i}\pi/3} - 1), \qquad X_{\pm 1}^C(0) = -\tfrac{1}{3} q \mathrm{e}^{\mp \mathrm{i}\pi/3}, \tag{2.22}$$

and these can be used to show that the rightmost sum in (2.21) approximately returns the voltages required in (2.15), so that

$$v_0^o(t) = v_{\mathrm{ref}}^o(t) + O(\Omega). \tag{2.23}$$

The $O(\Omega)$ error results from the discrete sampling of the input voltages used to compute the modulation strategy with regular sampling. It can be eliminated by changing the modulation strategy to natural sampling, which is now described.

2.4. A model calculation for the case of natural sampling. We now turn to the case of natural sampling. Here, in contrast to regular sampling, the input voltages are monitored continuously, and switching takes place at the instants when these voltages become equal to some other reference voltage. Because of the need to continuously monitor the input voltages, such sampling is generally implemented using analogue electronics. The key additional algebraic complication associated with natural sampling is that the switching times satisfy *implicit* equations. To analyze *natural sampling*, then, we consider once again the model sum in (2.5) except that the switching times are chosen to satisfy conditions

$$\alpha_n = A(n + \alpha_n) \qquad \text{and} \qquad \beta_n = B(n + \beta_n), \tag{2.24}$$

where $A(t) = \mu_\alpha + \lambda_\alpha \cos(\Omega t + \delta_0)$ and $B(t) = \mu_\beta + \lambda_\beta \cos(\Omega t + \delta_0 + \delta_1)$.

We note that, according to (2.24), we may consider α_n and β_n to be *irregular* samples of the continuous functions $A(t)$ and $B(t)$. However, to make analytical headway with our approach, it is preferable instead to regard α_n and β_n as being obtained by *regularly* sampling continuous functions $\alpha(\tau)$ and $\beta(\tau)$ as in (2.7). Now, however, it is the functions $A(t)$ and $B(t)$ that are prescribed explicitly while the functions $\alpha(\tau)$ and $\beta(\tau)$ are determined implicitly by

$$\alpha(\tau) = A(\tau + \alpha(\tau)) \qquad \text{and} \qquad \beta(\tau) = B(\tau + \beta(\tau)),$$

which are continuous versions of (2.24).

In the case of natural sampling, the output spectrum is best calculated by performing Poisson resummation *before* taking the Fourier transform of the function $F(t)$. Using the version

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{2\pi n \mathrm{i}\tau} f(\tau)\,\mathrm{d}\tau \tag{2.25}$$

of Poisson resummation on expression (2.5) for $F(t)$, we find that

$$F(t) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{2\pi n \mathrm{i}\tau} \psi_{\tau+\alpha(\tau),\tau+\beta(\tau)}(t)\,\mathrm{d}\tau = \sum_{n=-\infty}^{\infty} \int_{\tau_B(t)}^{\tau_A(t)} \mathrm{e}^{2\pi n \mathrm{i}\tau}\,\mathrm{d}\tau, \tag{2.26}$$

where $\tau_B(t)$ and $\tau_A(t)$ are, respectively, the values of τ at which the step function $\psi_{\tau+\alpha(\tau),\tau+\beta(\tau)}(t)$ switches on and then off again for fixed t. These switching times satisfy the conditions $t = \tau_A + \alpha(\tau_A)$ and $t = \tau_B + \beta(\tau_B)$, which can be rearranged to give

$$\tau_A(t) = t - \alpha(\tau_A) = t - A(t) \qquad \text{and} \qquad \tau_B(t) = t - \beta(\tau_B) = t - B(t).$$

We therefore have

$$F(t) = \sum_{n=-\infty}^{\infty} \int_{t-B(t)}^{t-A(t)} \mathrm{e}^{2\pi n \mathrm{i}\tau}\,\mathrm{d}\tau = \sum_{n=-\infty}^{\infty} \mathrm{e}^{2\pi n \mathrm{i}t} F_n(t), \tag{2.27}$$

where

$$F_n(t) = \int_{-B(t)}^{-A(t)} \mathrm{e}^{2\pi n \mathrm{i}\tau}\,\mathrm{d}\tau = \begin{cases} B(t) - A(t) & \text{if } n = 0, \\ \dfrac{\mathrm{e}^{-2\pi n \mathrm{i} B(t)} - \mathrm{e}^{-2\pi n \mathrm{i} A(t)}}{-2\pi n \mathrm{i}} & \text{otherwise.} \end{cases} \tag{2.28}$$

Equation (2.8) now gives

$$F(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m(2\pi n)\,\mathrm{e}^{\mathrm{i}\omega_{nm}t + \mathrm{i}m\delta^o},$$

where $X_m(\omega)$ has been defined in (2.9) and, in the special case $n = 0$, we may use (2.10); ω_{nm} has been defined in (2.12). The Fourier transform is

$$\hat{F}(\omega) = 2\pi \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} X_m(2\pi n)\,\mathrm{e}^{m\mathrm{i}\delta^o} \delta(\omega - \omega_{nm}). \tag{2.29}$$

These results are similar to those given in section 2.2 for the case of uniform sampling, except that the amplitude functions X_m are evaluated at different values of the argument ($2\pi n$ here, rather than ω_{nm} for uniform sampling). This difference has a dramatic effect on the terms with $n = 0$, however, which collectively contribute

$$F_0(t) = \sum_{m=-\infty}^{\infty} X_m(0)\,\mathrm{e}^{\mathrm{i}\omega_{0m}t + \mathrm{i}m\delta^o} = B(t) - A(t)$$

to the sum (see also (2.28)). This is the natural-sampling analogue of (2.14) and it is *exact*, which means that the *only* low frequencies present in $F(t)$ are those present in the prescribed functions $A(t)$ and $B(t)$ (cf. [3, 6, 8]).

2.5. Output voltages in the case of natural sampling. We now adapt the model calculation in section 2.4 to the case of three-phase output voltages produced by naturally sampling input voltages as in (2.2). The difference from the calculation in section 2.3 is that here the switching times are determined implicitly by equations of the form

$$\alpha_n^o = A^o(n+\alpha_n^o) \qquad \text{and} \qquad \beta_n^o = B^o(n+\beta_n^o),$$

where

$$A^o(\tau) = \tfrac{1}{3} + \tfrac{2}{3}q\cos(\Omega\tau+\delta^o), \qquad B^o(\tau) = \tfrac{2}{3} + \tfrac{2}{3}q\cos(\Omega\tau+\delta^o-\tfrac{\pi}{3}), \tag{2.30}$$

rather than being given directly as in (2.16).

There is nothing fundamentally new in the calculation here that has not already been covered in sections 2.3 and 2.4 so we simply present the main results. The output voltages can be given in the form

$$v^o(t) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \tilde{V}_{nm}\mathrm{e}^{i\Omega_{nm}t+\mathrm{i}m\delta^o}, \tag{2.31}$$

which is similar to (2.20), except that in the expression

$$\tilde{V}_{nm} = X_m^A(2\pi n) + pX_m^B(2\pi n) + p^2X_m^C(2\pi n),$$

the arguments of the functions X_m^i (which are once again given by (2.19)) are $2\pi n$ instead of ω_{nm}. The part of $v^o(t)$ of most physical interest is the contribution from terms with $n=0$; this contribution can be written, using (2.10),

$$v_0^o(t) = \sum_{m=-1}^{1}\left(X_m^A(0) + pX_m^B(0) + p^2X_m^C(0)\right)\mathrm{e}^{i\Omega_{0m}t+\mathrm{i}m\delta^o} = q\mathrm{e}^{\mathrm{i}\omega_1 t+\mathrm{i}\delta^o}. \tag{2.32}$$

Remarkably, this result coincides exactly with the desired form in (2.15). We emphasize that the result $v_0^o(t) = v_{\mathrm{ref}}^o(t)$ is *exact* for natural sampling (cf. the corresponding result (2.23) for uniform sampling, where there are errors of order Ω). In other contexts, this exact capture of some reference output is well known [3, 6, 7, 8].

3. Input and output currents. We now turn our attention to the currents in the system. The *output* currents are readily determined from the output voltages, provided that the output impedances are known (for simplicity, we suppose that the output neutral is connected to the supply neutral—see Figure 2.1). In order to construct the *input* currents, however, we must examine how the modulation strategy assigns the output currents to each input line; we are thus led to consider *two* separate discrete sampling processes, and the Fourier transform is as a result more complex to analyze. We note that in practice the input currents are monitored to provide a diagnostic of the system, and thus a knowledge of their spectrum is of particular practical utility.

The general discussion below applies equally to either regular or natural sampling. In section 3.4 below, we specialize the analysis to the two cases separately.

3.1. Notation for currents. We begin by setting out notation, building on the discussion in section 2.1. We adopt similar conventions for the input and output currents, writing

$$\mathbf{i}^{\text{in}}(t) = (i^A, i^B, i^C)^T \quad \text{and} \quad \mathbf{i}^{\text{out}}(t) = (i^a, i^b, i^c)^T,$$

where the superscript T denotes the transpose, and we note that current conservation means that these are connected by the transpose of the switching matrix [12], according to

$$\mathbf{i}^{\text{in}}(t) = M(t)^T \mathbf{i}^{\text{out}}(t). \tag{3.1}$$

A central goal in this paper is to compute the spectrum of the input currents in terms of the input voltages, and there are two elements to this calculation: summing over the windows of time in which the input voltage is sampled as a simple harmonic, and calculating the contributions from individual windows within that sum.

The summation is a double sum arising from the combined matrix products in (2.3) and (3.1) and is described explicitly below. We first describe in general terms the contribution of an individual element in this sum.

3.2. Loading the output: Currents associated with individual input pulses. Let

$$v_{t_3,t_4}(t) = \psi_{t_3,t_4}(t)\mathrm{e}^{i\omega_0 t} \tag{3.2}$$

represent an output voltage obtained by sampling a harmonic input voltage $\mathrm{e}^{i\omega_0 t}$ over the window $t_3 < t < t_4$. Let the output be connected to a load described by the impedance $Z(\omega)$, so that in the frequency domain the output current is

$$\hat{\imath}_{t_3,t_4}(\omega) = \frac{1}{Z(\omega)}\hat{v}_{t_3,t_4}(\omega) = \frac{1}{Z(\omega)}\hat{\psi}_{t_3,t_4}(\omega - \omega_0).$$

We describe the corresponding relation in the time domain using an admittance operator Y, such that $i_{t_3,t_4}(t) = Yv_{t_3,t_4}(t)$. This will describe an output current that switches on at $t = t_3$, is driven harmonically in the window $t_3 < t < t_4$, and decays as a transient thereafter, when $t_4 < t < \infty$. If a given input line connects to the output in question during the window $t_1 < t < t_2$, then we denote the corresponding contribution to that input current by

$$i_{t_1,t_2,t_3,t_4}(t) = \psi_{t_1,t_2}(t) i_{t_3,t_4}(t) = \psi_{t_1,t_2}(t) Y v_{t_3,t_4}(t). \tag{3.3}$$

The corresponding relation in the frequency domain is

$$\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega) = \frac{1}{2\pi}\hat{\psi}_{t_1,t_2}(\omega) * \hat{\imath}_{t_3,t_4}(\omega) = \frac{1}{2\pi}\hat{\psi}_{t_1,t_2}(\omega) * \left[\frac{1}{Z(\omega)}\hat{v}_{t_3,t_4}(\omega)\right], \tag{3.4}$$

where $*$ denotes convolution.

3.3. Loading the output: Total currents. Net input currents are obtained by summing individual contributions of the form (3.4), as governed by the matrix products in (2.3) and (3.1). We now outline details and notation for this process. Let the output voltages and currents be related by

$$\mathbf{i}^{\text{out}}(t) = \mathcal{Y}\mathbf{v}^{\text{out}}(t), \tag{3.5}$$

where $\mathcal{Y}$ is the diagonal matrix of admittance operators

$$\mathcal{Y} = \begin{pmatrix} Y^a & 0 & 0 \\ 0 & Y^b & 0 \\ 0 & 0 & Y^c \end{pmatrix} \equiv \operatorname{diag}(Y^a, Y^b, Y^c),$$

and where the diagonal elements are specific to each output. In Fourier representation, the admittance operator is represented by the simple diagonal matrix

$$\mathcal{Y}(\omega) = \mathrm{diag}(Z^a(\omega)^{-1}, Z^b(\omega)^{-1}, Z^c(\omega)^{-1})$$

of output-specific admittances. Combining (3.5) with (2.3) and (3.1), we may write

$$\mathbf{i}^{\mathrm{in}}(t) = M(t)^T \mathcal{Y} M(t) \mathbf{v}^{\mathrm{in}}(t).$$

Let us denote by

$$Q(t) = M(t)^T \mathcal{Y} M(t) \mathrm{e}^{i\omega_0 t}$$

the combined operator relating the input currents to the (known) input voltages, so that

$$\mathbf{i}^{\mathrm{in}}(t) = Q(t) \begin{pmatrix} 1 \\ p \\ p^2 \end{pmatrix}. \tag{3.6}$$

A typical element of $Q(t)$ can be written as a simple sum over outputs. Specifically,

$$Q^{ij}(t) = \sum_{o=a,b,c} F^{oi}(t) Y^o F^{oj}(t) \mathrm{e}^{i\omega_0 t}, \tag{3.7}$$

where the row index $i = A, B, C$ and the column index $j = A, B, C$ are labels of inputs. Each switching element $F^{oi}(t)$ is, in fact, a train of step-functions, as illustrated in (2.4). This allows us to write more explicitly, for example,

$$Q^{BB}(t) = \sum_{o=a,b,c} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \psi_{m+\alpha_m^o, m+\beta_m^o}(t) Y^o \psi_{n+\alpha_n^o, n+\beta_n^o}(t) \mathrm{e}^{i\omega_0 t}. \tag{3.8}$$

Note that the individual terms in this sum are of the form given in (3.3), with $t_1 = m + \alpha_m^o$, $t_2 = m + \beta_m^o$, $t_3 = n + \alpha_n^o$, and $t_4 = n + \beta_n^o$. Other entries in the matrix $Q(t)$ can be written similarly, except that alternative combinations of switching times are substituted for t_1, t_2, t_3, and t_4.

In the frequency domain, a typical element of the matrix $\hat{Q}(\omega)$ can be written, in analogy with (3.7),

$$\hat{Q}^{ij}(\omega) = \frac{1}{2\pi} \sum_{o=a,b,c} \hat{F}^{oi}(\omega) * \left[\frac{1}{Z^o(\omega)} \hat{F}^{oj}(\omega - \omega_0) \right]. \tag{3.9}$$

3.4. Direct calculation of input currents. We now outline a direct calculation of the input currents, for regular or natural sampling, using (3.8) and (3.9), respectively, as a basis. Nothing is assumed here about the form of the output impedances and the method is very general. More efficient, but less general, methods are described later, for specific forms of the output impedances. The difference between the two sets of calculations derives from whether we perform the convolution integral in (3.9) before or after the double sum over switching times. In this section, the sum is performed first and the convolution after.

3.4.1. Regular sampling. We first describe the calculation for regular sampling. For the contribution $\hat{Q}^{ij}(\omega)$, as described by (3.9), we first write

$$\hat{Q}^{ij}(\omega) = \sum_{o=a,b,c} \hat{Q}^{o,ij}(\omega),$$

where

$$\hat{Q}^{o,ij}(\omega) = \frac{1}{2\pi}\hat{F}^{oi}(\omega) * \left[\frac{1}{Z^o(\omega)}\hat{F}^{oj}(\omega - \omega_0)\right], \tag{3.10}$$

and the function $\hat{F}^{oi}(\omega)$ has been defined in (2.17). In this expression the function

$$\hat{F}^{oi}(\omega) = 2\pi \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} X_m^i(\omega_{nm})\mathrm{e}^{m\mathrm{i}\delta^o}\delta(\omega - \omega_{mn})$$

is convolved with

$$\frac{1}{Z^o(\omega)}\hat{F}^{oj}(\omega - \omega_0) = 2\pi \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} \frac{X_m^j(\omega_{nm})\mathrm{e}^{m\mathrm{i}\delta^o}}{Z^o(\omega_0 + \omega_{mn})}\delta(\omega - \omega_0 - \omega_{mn}),$$

and the result is a quadruple sum

$$\hat{Q}^{o,ij}(\omega) = 2\pi \sum_{klnm} W_{klnm}^{o,ij}\delta(\omega - \omega_0 - \omega_{kl} - \omega_{nm}),$$

where

$$W_{klnm}^{o,ij} = X_l^i(\omega_{kl})X_m^j(\omega_{nm})\left[\frac{\mathrm{e}^{\mathrm{i}(l+m)\delta^o}}{Z^o(\omega_0 + \omega_{nm})}\right]$$

and, in the sum, the indices k, l, n, and m run independently from $-\infty$ to ∞. The output-dependent parts (to be summed over later) have been isolated within square brackets in this expression. Using the fact that $\omega_{kl} + \omega_{nm} = \omega_{k+n,l+m}$, this result can alternatively be stated in the time domain as

$$Q^{o,ij}(t) = \sum_{klnm} W_{klnm}^{o,ij}\mathrm{e}^{\mathrm{i}(\omega_0 + \omega_{k+n,l+m})t}.$$

It is convenient to group terms in this sum with a common frequency, giving

$$Q^{o,ij}(t) = \sum_{NM} \mathcal{Q}_{NM}^{o,ij}\mathrm{e}^{\mathrm{i}\Omega_{NM}t}, \tag{3.11}$$

where $\Omega_{NM} = \omega_0 + \omega_{NM} = \omega_0 + 2\pi N + M\Omega$ and

$$\mathcal{Q}_{NM}^{o,ij} = \sum_{nm} W_{N-n,M-m,n,m}^{o,ij}. \tag{3.12}$$

In this result, the amplitude $\mathcal{Q}_{NM}^{o,ij}$ of a term with a given frequency Ω_{NM} is expressed as a double sum. Finally, we note that, according to (3.6), the total current in an input labeled by the superscript i can be obtained from the results above using

$$i^i(t) = \sum_{o=a,b,c} Q^{o,iA}(t) + pQ^{o,iB}(t) + p^2Q^{o,iC}(t). \tag{3.13}$$

Although these formulas for the input currents seem rather unwieldy, we may already note the potentially diagnostically useful result that, if all outputs have equal impedance, then the sum over outputs produces a factor $\sum_o \mathrm{e}^{M\mathrm{i}\delta^o} = 1 + \mathrm{e}^{2\pi M\mathrm{i}/3} + \mathrm{e}^{4\pi M\mathrm{i}/3}$, which vanishes unless M is a multiple of 3. Hence, in this special case, the frequencies $\Omega_{NM} = \omega_0 + 2\pi N + M\Omega$ appear only where M is a multiple of 3.

3.4.2. Natural sampling. We may readily adapt the expressions derived in the previous section to the case of natural sampling, without repeating the entire calculation. To do so, we simply note, from (2.18) and (2.29), that the only material difference in the natural sampling case lies in the arguments of the functions X_m^i used in the definitions of $\hat{F}^{oi}(\omega)$. Hence the input currents are still given by an expression of the form (3.13), with $Q^{o,ij}$ given by (3.11); however, $\mathcal{Q}_{NM}^{o,ij}$ is now given by

$$\mathcal{Q}_{NM}^{o,ij} = \sum_{nm} X_{M-m}^i(2\pi(N-n))X_m^j(2\pi n)\left[\frac{\mathrm{e}^{\mathrm{i}M\delta^o}}{Z^o(\omega_0+\omega_{nm})}\right], \tag{3.14}$$

rather than by (3.12). It is interesting to note that in the case of input currents, unlike in the cases of output voltages and currents, natural sampling does not produce a clean single harmonic when the high-frequency terms with $N \neq 0$ are filtered out.

4. Results for the voltage and current spectra. In this section we illustrate the results above for the output voltage and input current spectra. Recalling that the switching frequency has been scaled to 2π, we choose parameter values as follows:

$$\omega_0 = \tfrac{1}{20} \times 2\pi, \qquad \omega_1 = \tfrac{1}{25} \times 2\pi, \qquad q = 0.4. \tag{4.1}$$

The corresponding output voltage spectrum is independent of the output loads, and is shown in Figure 4.1, for regular and natural sampling. The primary feature of note is the exact reproduction of the low-frequency ($n = 0$) part of the voltage spectrum for natural sampling and, in contrast, the significant low-frequency distortion introduced by regular sampling.

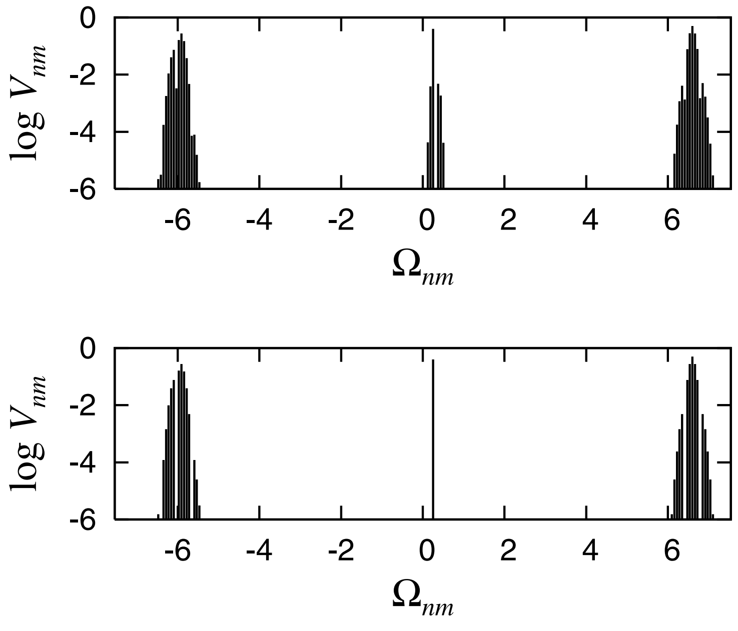


FIG. 4.1. *Spectrum of the output voltages, for parameter values* (4.1). *Upper plot: regular sampling. Lower plot: natural sampling. Note the significant low-frequency distortion of the regularly sampled case.*

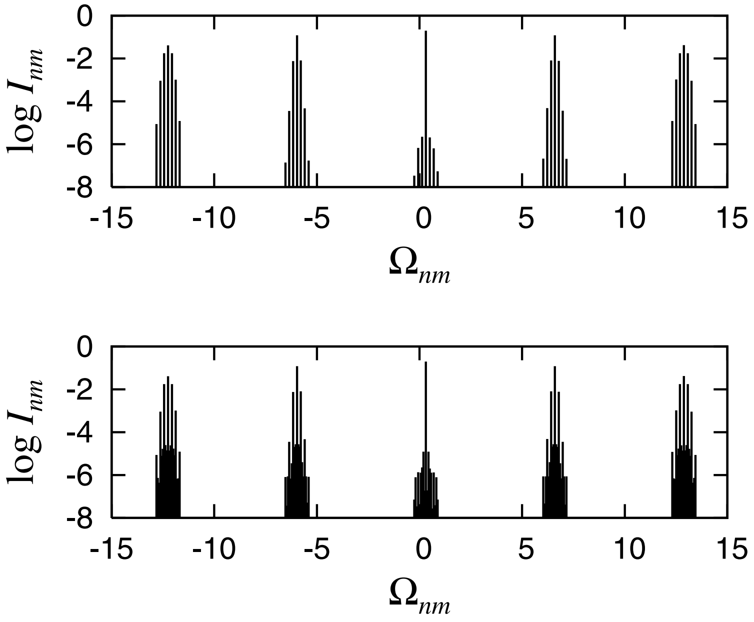


FIG. 4.2. *Spectrum of the input currents in line A with regular sampling, for parameter values* (4.1). *Upper plot: balanced loads, with* (4.3). *Lower plot: unbalanced loads, with* (4.4). *Spectra for input lines B and C are similar.*

To determine the input currents, we must specify the output loads, which will each comprise a resistor and an inductor in series, so that

$$Z^o(\omega) = R^o + \mathrm{i}\omega L^o, \quad o = a,\, b, \text{ or } c. \tag{4.2}$$

We consider two cases. In the first, all output lines offer equal impedance, with

$$R^a = R^b = R^c = 5\Omega, \qquad L^a = L^b = L^c = 5\text{mH}; \tag{4.3}$$

in the second, the output impedances are unbalanced, and we take

$$R^a = R^b = R^c = 5\Omega, \qquad L^a = L^b = 5\text{mH}, \quad L^c = 0\text{mH}. \tag{4.4}$$

From Figure 4.2 we see that the frequency spectrum for the input current is sparser for the balanced load. In fact, as argued above, the spectrum is confined to frequencies of the form $\omega_0 + 2\pi n + m\Omega$, where m is a multiple of 3; for the unbalanced load, by contrast, all frequencies of the form $\omega_0 + 2\pi n + m\Omega$ are present.

5. More efficient calculation of the input spectrum. The results of section 3 provide expressions for the input and output currents, and thus solve the problem posed at the start of this paper. However, each coefficient in the input current spectrum requires the evaluation of a doubly infinite sum, as in (3.12) and (3.14). Furthermore, these sums converge rather slowly. So we now describe a more efficient means of obtaining the input current spectrum for specific output impedances. It differs from the calculation in section 3 by taking advantage of the known impedances to perform the convolution integrals in (3.9) before the sum over switching times is evaluated. A consequence is that each Fourier coefficient will then require calculation of only a single infinite sum.

The explicit calculations outlined are for the case where each output load takes the form of a resistor and an inductor in series, so that the corresponding impedances are given by (4.2). The various complex impedances need not be equal between outputs a, b, and c. We shall discuss later how this method may be extended to the case of more general forms for the output impedances.

5.1. Illustration of transfer matrix calculation for purely resistive loads. We begin by considering purely resistive loads, with $L^o = 0$. Although such loads are of limited practical interest, they are nevertheless useful to illustrate the following transfer matrix method. For a purely resistive load, the admittances of the three output lines are simply the constants $Y^o = 1/R^o$. This case is thus considerably easier to analyze than that of general impedance because a step output voltage of the sort described in section 3.2 produces an output current only while the voltage is switched on. Then, for example, the terms in the sum (3.7) vanish unless $n = m$ and we find that the diagonal terms in Q take the form

$$Q^{ii}(t) = \sum_{o=a,b,c} \frac{1}{R^o} F^{oi}(t) \mathrm{e}^{\mathrm{i}\omega_0 t}, \quad i = A,\, B, \text{ or } C. \tag{5.1}$$

Note that the functions $F^{oi}(t)$ have been defined in (2.17). The off-diagonal terms in Q necessarily vanish for resistive loads. To see this, consider, for example,

$$Q^{AB}(t) = \sum_{o=a,b,c} \sum_{m,n} \frac{1}{R^o} \psi_{m,m+\alpha_m^o}(t) \psi_{n+\alpha_n^o, n+\beta_n^o}(t). \tag{5.2}$$

It is clear that the intervals $(m, m+\alpha_m^o)$ and $(n+\alpha_n^o,\, n+\beta_n^o)$ never overlap and hence $Q^{AB} = 0$. A similar consideration shows that all other off-diagonal terms are zero.

The three input currents are then, using the results and notation of sections 2.3 and 2.5,

$$\begin{aligned} i^A(t) = Q^{AA}(t) \quad &= \sum_{nm} Y_m X_m^A(x_{nm}) \mathrm{e}^{\mathrm{i}\Omega_{nm} t}, \\ i^B(t) = pQ^{BB}(t) \quad &= \sum_{nm} pY_m X_m^B(x_{nm}) \mathrm{e}^{\mathrm{i}\Omega_{nm} t}, \\ i^C(t) = p^2 Q^{CC}(t) &= \sum_{nm} p^2 Y_m X_m^C(x_{nm}) \mathrm{e}^{\mathrm{i}\Omega_{nm} t}, \end{aligned}$$

where $x_{nm} = \omega_{nm}$ for regular sampling, and $x_{nm} = 2\pi n$ for natural sampling. Here

$$Y_m = \sum_{o=a,b,c} \frac{\mathrm{e}^{\mathrm{i}m\delta^o}}{R^o}$$

is an effective total output admittance, common to all three inputs.

These two simplifying elements of the matrix Q (namely that Q is a *diagonal* matrix, and that the diagonal elements are given by a *single* sum, as in (5.1)) follow from there being a purely resistive load. The key points are that an output voltage pulse produces a proportional output current pulse, and that after the voltage pulse the corresponding current drops immediately to zero.

5.2. Frequency domain calculation for series resistor/inductor loads. We now consider the case of more general loads, with inductive as well as resistive elements, with output impedances given by (4.2). In this case, the double sum in (3.9) does not collapse to a single sum as it did in the purely resistive case and the calculation is more complex.

Before summing the series, let us consider in more detail the structure of the summand, whose general form is given in (3.4). For an inductive load with impedance $Z = R + i\omega L = iL(\omega - i\gamma)$, where $\gamma = R/L$ is the decay rate of transients in the current, we can write (3.4) more explicitly as

$$\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega) = \frac{1}{2\pi \mathrm{i}} \int_{-\infty}^{\infty} \frac{(\mathrm{e}^{-\mathrm{i}(\omega-\omega')t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\omega')t_1})(\mathrm{e}^{-\mathrm{i}(\omega'-\omega_0)t_4} - \mathrm{e}^{-\mathrm{i}(\omega'-\omega_0)t_3})}{L(\omega'-\omega)(\omega'-\mathrm{i}\gamma)(\omega'-\omega_0)}\, \mathrm{d}\omega',$$

having substituted the explicit forms for $\hat{\psi}_{t_1,t_2}(\omega)$ and $\hat{\psi}_{t_3,t_4}(\omega)$ in the convolution integral. This is conveniently evaluated using the residue calculus. The denominator in the integrand has three zeroes, two on the real axis and one on the positive imaginary axis. The zeroes on the real axis are canceled by zeroes in the numerator and do not lead to poles in the total integrand. Since the integrand is analytic on the real axis, we may move the contour slightly off the real axis before beginning the calculation proper and the manner in which we do this will not affect the final result. This observation is relevant because we will evaluate the integral by expanding the numerator and considering terms individually. Although these individual terms have poles on the real axis, if we have deformed the contour away from these poles beforehand, the individual integrals are well defined. Furthermore, any contributions made by the poles on the real axis through the residue calculus must combine consistently and be independent of the initial contour deformation. The integral is therefore controlled by the pole at $\omega' = i\gamma$. There are three cases to consider.

Case 1. *Window t_1, t_2 precedes window t_3, t_4.*
In this case, where $t_1 < t_2 < t_3 < t_4$, every term in the expanded numerator

$$\begin{aligned}(\mathrm{e}^{-\mathrm{i}(\omega-\omega')t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\omega')t_1})&(\mathrm{e}^{-\mathrm{i}(\omega'-\omega_0)t_4} - \mathrm{e}^{-\mathrm{i}(\omega'-\omega_0)t_3}) \\ &= \mathrm{e}^{-\mathrm{i}\omega t_2 + \mathrm{i}\omega_0 t_4 + \mathrm{i}\omega'(t_2-t_4)} - \mathrm{e}^{-\mathrm{i}\omega t_2 + \mathrm{i}\omega_0 t_3 + \mathrm{i}\omega'(t_2-t_3)} \\ &\quad - \mathrm{e}^{-\mathrm{i}\omega t_1 + \mathrm{i}\omega_0 t_4 + \mathrm{i}\omega'(t_1-t_4)} + \mathrm{e}^{-\mathrm{i}\omega t_1 + \mathrm{i}\omega_0 t_3 + \mathrm{i}\omega'(t_1-t_3)}\end{aligned} \tag{5.3}$$

decays exponentially as ω' descends into the lower-half plane. The contour of integration can therefore be pushed downwards and, because there are no poles in the lower-half plane, the integral must vanish. We therefore necessarily have

$$\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega) = 0 \tag{5.4}$$

in this case. Note that in the case of more complicated load impedances, causality demands that all of the zeroes of $Z(\omega')$ lie in the upper-half plane and the result (5.4) still holds. The result is obvious in the time domain because a driving voltage confined to the window (t_3, t_4) produces no current for $t < t_3$, and any sampling window confined to this range must produce a null result.

Case 2. *Window t_1, t_2 follows window t_3, t_4.*
In this case, $t_3 < t_4 < t_1 < t_2$. Let us assume that the integration contour has been moved slightly above the real axis before the calculation for individual terms in expansion (5.3) begins. Then, all of the terms on the right of (5.3) are exponentially

decaying into the upper-half plane and we push the contour upwards, picking up a contribution only from the pole at $\omega' = \mathrm{i}\gamma$. This yields

$$\begin{aligned}\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega) &= \frac{(\mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_1})(\mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_4} - \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_3})}{L(\omega-\mathrm{i}\gamma)(\omega_0-\mathrm{i}\gamma)}\\ &= \frac{1}{L}\left(\frac{\mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_1}}{-\mathrm{i}(\omega-\mathrm{i}\gamma)}\right)\left(\frac{\mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_4} - \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_3}}{\mathrm{i}(\omega_0-\mathrm{i}\gamma)}\right)\\ &= \frac{1}{L}\hat{\psi}_{t_1,t_2}(\omega-\mathrm{i}\gamma)\hat{\psi}_{t_3,t_4}(-\omega_0+\mathrm{i}\gamma). \qquad (5.5)\end{aligned}$$

For more general load impedances, there would be a sum of such contributions, each term corresponding to a zero of $Z(\omega')$, or equivalently a decay rate of the system, weighted by residues of $1/Z(\omega')$.

Case 3. *Window t_1, t_2 coincides with window t_3, t_4.*
In this case, $t_1 = t_3 < t_2 = t_4$. The calculation here is somewhat more complicated and requires a more careful consideration of the poles on the real axis. We forgo the details here and simply quote the result:

$$\begin{aligned}\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega) &= \frac{1}{\mathrm{i}L(\omega_0-\mathrm{i}\gamma)}\left(\frac{\mathrm{e}^{-\mathrm{i}(\omega-\omega_0)t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\omega_0)t_1}}{-\mathrm{i}(\omega-\omega_0)}\right)\\ &\qquad - \frac{\mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_1}}{\mathrm{i}L(\omega_0-\mathrm{i}\gamma)}\left(\frac{\mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_2} - \mathrm{e}^{-\mathrm{i}(\omega-\mathrm{i}\gamma)t_1}}{-\mathrm{i}(\omega-\mathrm{i}\gamma)}\right)\\ &= \frac{1}{Z(\omega_0)}\left[\hat{\psi}_{t_1,t_2}(\omega-\omega_0) - \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma)t_1}\hat{\psi}_{t_1,t_2}(\omega-\mathrm{i}\gamma)\right], \qquad (5.6)\end{aligned}$$

which is obtained by summing the contributions from the various poles for each of the terms in (5.3) and following some further algebraic manipulation.

So far we have established the forms of the individual terms in (3.9). It remains to perform the double sum over switching times in that equation. Once again we concentrate initially on the element $\hat{Q}^{BB}(\omega)$ and indicate later how the calculation is altered for other elements. In this case, the summands in (3.9) are of the form $\hat{\imath}_{t_1,t_2,t_3,t_4}(\omega)$, as calculated above, with $t_1 = m + \alpha^o_m$, $t_2 = m + \beta^o_m$, $t_3 = n + \alpha^o_n$, $t_4 = n + \beta^o_n$, and with loads that are output-specific. Let us denote by

$$\gamma^o = R^o/L^o \qquad (5.7)$$

the decay rate of transients associated with output o. In view of (5.4), the summands vanish if $n < m$, so let us set $n = m + r$ for $r = 0, 1, 2, \ldots$, and separate (3.9) into "diagonal" and "off-diagonal" contributions:

$$\hat{Q}^{BB}(\omega) = \hat{Q}^{BB}_{\mathrm{diag}}(\omega) + \hat{Q}^{BB}_{\mathrm{offdiag}}(\omega),$$

where

$$\hat{Q}^{BB}_{\mathrm{diag}}(\omega) = \sum_{o=a,b,c}\sum_{n=-\infty}^{\infty}\hat{\imath}_{n+\alpha^o_n,n+\beta^o_n,n+\alpha^o_n,n+\beta^o_n}(\omega)$$

and

$$\hat{Q}^{BB}_{\mathrm{offdiag}}(\omega) = \sum_{o=a,b,c}\sum_{r=1}^{\infty}\sum_{n=-\infty}^{\infty}\hat{\imath}_{n+r+\alpha^o_{n+r},n+r+\beta^o_{n+r},n+\alpha^o_n,n+\beta^o_n}(\omega).$$

The diagonal contribution accounts for the terms with $r = 0$ and corresponds to the case of coinciding windows given in (5.6). We expect these terms to dominate the total sum and so describe them first. We start with regular sampling (natural sampling is discussed at the end of this section). Then using the notation of section 2.3, we write

$$\begin{aligned}\sum_{n=-\infty}^{\infty} \hat{\psi}_{t_1,t_2}(\omega-\omega_0) &= \sum_{n=-\infty}^{\infty} \hat{\psi}_{n+\alpha_n^o,n+\beta_n^o}(\omega-\omega_0) = \hat{F}^{oB}(\omega-\omega_0)\\ &= 2\pi \sum_{nm} X_m^B(\omega_{nm}) \mathrm{e}^{mi\delta^o} \delta(\omega-\Omega_{nm}),\end{aligned}$$

recalling that $\Omega_{nm} = \omega_0 + \omega_{nm} = \omega_0 + 2\pi n + m\Omega$. A similar calculation shows that

$$\begin{aligned}\sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma^o)t_1} \hat{\psi}_{t_1,t_2}(\omega-\mathrm{i}\gamma^o) &= \sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma^o)(n+\alpha_n^o)} \hat{\psi}_{n+\alpha_n^o,n+\beta_n^o}(\omega-\mathrm{i}\gamma^o)\\ &= \sum_{n=-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}n(\omega-\omega_0)}\, \mathrm{e}^{\mathrm{i}(\omega_0-\mathrm{i}\gamma^o)\alpha_n^o}\, \hat{\psi}_{\alpha_n^o,\beta_n^o}(\omega-\mathrm{i}\gamma^o)\\ &= 2\pi \sum_{nm} Y_m^B(\omega_{nm}, -\omega_0+\mathrm{i}\gamma^o) \mathrm{e}^{\mathrm{i}m\delta^o} \delta(\omega-\Omega_{nm}),\end{aligned}$$

where

$$Y_m^B(\omega,\omega') \equiv \sum_{k=-\infty}^{\infty} X_k^B(\omega-\omega') C_{m-k}^B(\omega') \tag{5.8}$$

and

$$C_m^B(\omega) = \mathrm{e}^{-\mathrm{i}\omega/3}(-\mathrm{i})^m \mathrm{J}_m\left(\tfrac{2}{3}q\omega\right).$$

Note that the expression defining $Y_m^B(\omega,\omega')$ can be summed using Graf's theorem [11]. However, leaving the definition of $Y_m^B(\omega,\omega')$ as a sum, as done here, has the advantage of admitting easier generalization to other diagonal terms and being simpler to write. We can therefore write

$$\hat{Q}^{BB}_{\mathrm{diag}}(\omega) = 2\pi \sum_{o=a,b,c} D_{nm}^{o,B} \mathrm{e}^{\mathrm{i}m\delta^o} \delta(\omega-\Omega_{nm}),$$

where

$$D_{nm}^{o,B} = \frac{1}{Z^o(\omega_0)} \left(X_m^B(\omega_{nm}) - Y_m^B(\omega_{nm}, -\omega_0+\mathrm{i}\gamma^o)\right).$$

There are similar diagonal contributions to $\hat{Q}^{AA}$ and $\hat{Q}^{CC}$, except that $C_k^A(\omega)$ and $C_k^C(\omega)$ have the alternative forms

$$C_m^A(\omega) = \delta_{m0}, \qquad C_m^C(\omega) = \mathrm{e}^{-2\mathrm{i}\omega/3-\mathrm{i}m\pi/3}(-\mathrm{i})^m \mathrm{J}_m\left(\tfrac{2}{3}q\omega\right).$$

We next discuss off-diagonal contributions. In this case the summands are of the form given in (5.5) and

$$
\begin{aligned}
\hat{Q}^{BB}_{\text{nondiag}}(\omega) &= \sum_{o=a,b,c} \sum_{r=1}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{L^o} \hat{\psi}_{t_1,t_2}(\omega - \mathrm{i}\gamma^o) \hat{\psi}_{t_3,t_4}(-\omega_0 + \mathrm{i}\gamma^o) \\
&= \sum_{o=a,b,c} \sum_{r=1}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{L^o} \mathrm{e}^{-\mathrm{i}n(\omega-\omega_0) - \mathrm{i}r(\omega - \mathrm{i}\gamma^o)} \\
&\qquad\qquad\qquad \times \hat{\psi}_{\alpha^o_{n+r},\beta^o_{n+r}}(\omega - \mathrm{i}\gamma^o) \hat{\psi}_{\alpha^o_n,\beta^o_n}(-\omega_0 + \mathrm{i}\gamma^o) \\
&= 2\pi \sum_{o=a,b,c} \sum_{nm} A^{o,BB}_{nm} \mathrm{e}^{\mathrm{i}m\delta^o} \delta(\omega - \Omega_{nm}), \qquad (5.9)
\end{aligned}
$$

where (after some manipulation)

$$
A^{o,BB}_{nm} = \frac{1}{L^o} U^{BB}_m(\omega_{nm}, -\omega_0 + \mathrm{i}\gamma^o)
$$

and

$$
U^{BB}_m(\omega, \omega') = \sum_{k=-\infty}^{\infty} G_1(\omega - \omega' - k\Omega) X^B_k(\omega - \omega') X^B_{m-k}(\omega')
$$

and

$$
G_1(\omega) = \sum_{r=1}^{\infty} \mathrm{e}^{-\mathrm{i}r\omega} = \frac{1}{\mathrm{e}^{\mathrm{i}\omega} - 1}.
$$

The off-diagonal contributions to $\hat{Q}^{AA}$ and $\hat{Q}^{CC}$ are of the same form, with appropriate replacements for X^B_m.

If we now consider elements $\hat{Q}^{ij}(\omega)$ with $i \neq j$, we find that the appropriate intervals (t_1, t_2) and (t_3, t_4) never overlap and all summands are of the form given in (5.5). The calculation is very similar to that for $\hat{Q}^{BB}_{\text{offdiag}}(\omega)$ except that, when $i > j$, the sum over r starts from $r = 0$ rather than $r = 1$ (here we adopt the convention that $C > B > A$). The result is the following generalization of (5.9):

$$
\hat{Q}^{ij}(\omega) = 2\pi \sum_{o=a,b,c} \sum_{nm} A^{o,ij}_{nm} \mathrm{e}^{\mathrm{i}m\delta^o} \delta(\omega - \Omega_{nm}), \qquad (5.10)
$$

where

$$
A^{o,ij}_{nm} = \frac{1}{L^o} U^{ij}_m(\omega_{nm}, -\omega_0 + \mathrm{i}\gamma^o)
$$

and

$$
U^{ij}_m(\omega, \omega') = \sum_{k=-\infty}^{\infty} G^{ij}(\omega - \omega' - k\Omega) X^i_k(\omega - \omega') X^j_{m-k}(\omega') \qquad (5.11)
$$

and

$$
G^{ij}(\omega) = \begin{cases} \mathrm{e}^{\mathrm{i}\omega} G_1(\omega) & \text{if } i > j, \\ G_1(\omega) & \text{if } i < j. \end{cases}
$$

Again, the sum defining $U^{ij}_m(\omega, \omega')$ can be expressed alternatively using Graf's theorem but the form given is simpler to write. Note that if we set $i = j$, then (5.10) also describes the off-diagonal part of $\hat{Q}^{ii}$, if we take $G^{ii}(\omega) = G_1(\omega)$.

The current in the input labeled by the subscript i is (cf. (3.13))

$$i^i(t) = Q^{iA}(t) + pQ^{iB}(t) + p^2 Q^{iC}(t), \tag{5.12}$$

where

$$Q^{ij}(t) = \sum_{o=a,b,c} \sum_{nm} \left(D^{o,i}_{nm} \delta_{ij} + A^{o,ij}_{nm} \right) \mathrm{e}^{\mathrm{i}m\delta^o} \mathrm{e}^{\mathrm{i}\Omega_{nm} t}.$$

Thus, from (5.12), the coefficient of each frequency component in the input current $i^i(t)$ requires (aside from sums over the three output lines) only a *single* infinite sum for computing each of the the terms $D^{o,i}_{nm}$ and $A^{o,ij}_{nm}$, as in (5.8) and (5.11).

Finally, we note also that a similar calculation is possible in the case of natural sampling. The answer in that case is similar, the main difference being that the functions X^i_m, Y^i_m, and U^{ij}_m take different arguments when they are used in the calculation of the amplitudes $D^{o,i}_{nm}$ and $A^{o,ij}_{nm}$ (compare (2.20) with (2.31), for example).

6. Conclusions. We have shown how to compute the output voltage spectrum, and the output and input current spectra for an idealized matrix power converter, for general output loads. Our method provides a rather more direct alternative to the usual approach of Black's multiple Fourier series [2, 3], and appears to be the first published calculation of the full spectrum. The mathematical expressions involved in the present calculations are considerably more compact than would be the equivalent expressions using Black's method. Despite its greater directness, however, our method still requires calculations that are rather algebraically involved. We have shown how reasonable assumptions about the form of the output loads—for example, if they are all series resistor-inductor loads—can be used for deriving more rapidly convergent expressions for the input currents (which are of particular significance since they provide an easily monitored diagnostic of the system). We note that the calculation in this paper can be adapted relatively easily to more general output impedances.

One potential practical upshot of our work is the following. In applications such as aeronautics, there are strict regulations regarding acceptable levels of the electromagnetic interference generated by high-frequency switching applications such as matrix converters. This paper provides, apparently for the first time, analytical expressions for the full frequency spectrum of voltages and currents. We therefore expect the formulas derived herein, and appropriate extensions of the methodology to more general cases (for example, a wider range of output impedances) to allow engineers to design matrix converters to satisfy mandatory restrictions on power quality without wasteful overspecification of the associated filters.

In the appendix, we illustrate how similar techniques can be adapted to more general switching protocols for the matrix converter. However, these introduce new frequencies into the spectrum, so the calculation is more involved.

The Fourier transform/Poisson resummation techniques applied here (with a judicious choice of the order in which the elements of the technique are applied, according to whether regular or natural sampling is used) may also be applied to other switching problems. Notable examples are the class-D audio amplifier, for which an analysis such as that given in this paper would lead to considerably more compact derivations of the spectrum than previously given [3, 8], and DC–AC converters (inverters) [7].

Appendix. The modulation strategies considered in the main text are the simplest possible; in practice, more complicated strategies are used. Many of these will be amenable to a treatment similar to that described in this paper, but with increased algebraic complexity. In this appendix we illustrate some of the necessary modifications by calculating the output voltage spectrum for natural sampling using a hybrid Venturini modulation strategy. In this case, the switching times are determined by

$$\begin{aligned}
A^o(\tau) &= \tfrac{1}{3} + \tfrac{2}{3}q\left[\theta\cos((\omega_1-\omega_0)\tau+\delta^o) + (1-\theta)\cos((\omega_1+\omega_0)\tau+\delta^o)\right],\\
B^o(\tau) &= \tfrac{2}{3} + \tfrac{2}{3}q\left[\theta\cos((\omega_1-\omega_0)\tau+\delta^o-\tfrac{\pi}{3}) + (1-\theta)\cos((\omega_1+\omega_0)\tau+\delta^o+\tfrac{\pi}{3})\right],
\end{aligned}$$

where $0 \le \theta \le 1$, rather than by (2.30). Notice that the case $\theta = 1$ recovers (2.30). The case $\theta = 1/2$ proves particularly straightforward to implement in practice [12].

The calculation of $F^{oi}(t)$ is now rather more involved, since

$$\mathrm{e}^{-2\pi n\mathrm{i}A^o(t)} = \mathrm{e}^{-2\pi n\mathrm{i}/3}\sum_{m=-\infty}^{\infty}\sum_{m'=-\infty}^{\infty}\mathcal{C}_{mm'}\mathrm{e}^{\mathrm{i}(m(\omega_1-\omega_0)+m'(\omega_1+\omega_0))t}$$

and

$$\mathrm{e}^{-2\pi n\mathrm{i}B^o(t)} = \mathrm{e}^{-4\pi n\mathrm{i}/3}\sum_{m=-\infty}^{\infty}\sum_{m'=-\infty}^{\infty}\mathcal{C}_{mm'}\mathrm{e}^{\mathrm{i}(m(\omega_1-\omega_0)+m'(\omega_1+\omega_0))t}\mathrm{e}^{\mathrm{i}(m'-m)\pi/3},$$

where

$$\mathcal{C}_{mm'} = (-\mathrm{i})^{m+m'}\mathrm{J}_m(\tfrac{4}{3}n\pi q\theta)\mathrm{J}_{m'}(\tfrac{4}{3}n\pi q(1-\theta))\mathrm{e}^{\mathrm{i}(m+m')\delta^o}.$$

Thus, in general, these quantities now involve additional frequencies beyond those present for the simpler case $\theta = 1$.

Writing the output voltages as

$$v^o(t) = \sum_{nmm'}\tilde{V}_{nmm'}\mathrm{e}^{\mathrm{i}(2\pi n+\omega_0+m(\omega_1-\omega_0)+m'(\omega_1+\omega_0))t}\mathrm{e}^{\mathrm{i}(m+m')\delta^o},$$

we have

$$\text{(A.1)}\qquad \tilde{V}_{nmm'} = X^A_{mm'}(2\pi n) + pX^B_{mm'}(2\pi n) + p^2X^C_{mm'}(2\pi n),$$

where

$$\begin{aligned}
X^A_{mm'}(\omega) &= \frac{(-\mathrm{i})^{m+m'}}{-\mathrm{i}\omega}\left[\mathrm{e}^{-\mathrm{i}\omega/3}\mathrm{J}_m(\tfrac{2}{3}q\omega\theta)\mathrm{J}_{m'}(\tfrac{2}{3}q\omega(1-\theta)) - \delta_{m0}\delta_{m'0}\right],\\
X^B_{mm'}(\omega) &= \frac{(-\mathrm{i})^{m+m'}}{-\mathrm{i}\omega}\left[\mathrm{e}^{-2\mathrm{i}\omega/3+\mathrm{i}(m'-m)\pi/3} - \mathrm{e}^{-\mathrm{i}\omega/3}\right]\mathrm{J}_m(\tfrac{2}{3}q\omega\theta)\mathrm{J}_{m'}(\tfrac{2}{3}q\omega(1-\theta)),\\
X^C_{mm'}(\omega) &= \frac{(-\mathrm{i})^{m+m'}}{-\mathrm{i}\omega}\left[\mathrm{e}^{-\mathrm{i}\omega}\delta_{m0}\delta_{m'0} - \mathrm{e}^{-2\mathrm{i}\omega/3+\mathrm{i}(m'-m)\pi/3}\mathrm{J}_m(\tfrac{2}{3}q\omega\theta)\mathrm{J}_{m'}(\tfrac{2}{3}q\omega(1-\theta))\right].
\end{aligned}$$

Special consideration needs to be given to the values of $X^i_{mm'}(0)$. We find that $X^i_{mm'}(0) = 0$ for $i = A$, B, or C, except in the following cases:

$$X^i_{00}(0) = \tfrac{1}{3}, \qquad i = A,\ B,\ \text{or}\ C$$

and

$$X^A_{0\pm1}(0) = \tfrac{1}{3}q(1-\theta), \quad X^B_{0\pm1}(0) = \tfrac{1}{3}q(1-\theta)(\mathrm{e}^{\pm\mathrm{i}\pi/3}-1), \quad X^C_{0\pm1}(0) = -\tfrac{1}{3}q(1-\theta)\mathrm{e}^{\pm\mathrm{i}\pi/3}$$

and

$$X^A_{\pm 10}(0) = \tfrac{1}{3}q\theta, \qquad X^B_{\pm 10}(0) = \tfrac{1}{3}q\theta(\mathrm{e}^{\mp \mathrm{i}\pi/3} - 1), \qquad X^C_{\pm 10}(0) = -\tfrac{1}{3}q\theta\mathrm{e}^{\mp \mathrm{i}\pi/3}.$$

It then follows from (A.1) that for the contribution to the output voltages with $n = 0$ we have $\tilde{V}_{00-1} = q(1-\theta)$ and $\tilde{V}_{010} = q\theta$, with $\tilde{V}_{0mm'} = 0$ for all other choices of m and m'. Thus the corresponding contribution to the output voltages is

$$v^o_0(t) = q\theta\mathrm{e}^{\mathrm{i}\omega_1 t + \mathrm{i}\delta^o} + q(1-\theta)\mathrm{e}^{-\mathrm{i}\omega_1 t - \mathrm{i}\delta^o}.$$

Hence the physical output voltage, given by the real part of this expression, is, as for the simpler modulation strategy of section 2.5, *exactly* the intended reference voltage $v^o_{\mathrm{ref}}(t)$. A similar calculation can be undertaken for uniform sampling but is not described here.

REFERENCES

[1] A. Alesina and M. G. B. Venturini, *Solid-state power conversion: A Fourier analysis approach to generalized transformer synthesis*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 319–330.

[2] W. R. Bennett, *New results in the calculation of modulation products*, Bell Syst. Tech. J., 12 (1933), pp. 228–243.

[3] H. S. Black, *Modulation Theory*, Van Nostrand, New York, 1953.

[4] S. R. Bowes, *New sinusoidal pulsewidth-modulated inverter*, Proc. IEEE, 122 (1975), pp. 1279–1285.

[5] J. C. Clare, L. Empringham, and P. W. Wheeler, *The effects of sampling delays and non-ideal filtering on the performance of matrix converter modulation algorithms*, in Proceedings of the Eighth International Conference on Power Electronics and Variable Speed Drives, London, 2000, pp. 29–34.

[6] S. M. Cox and B. H. Candy, *Class-D audio amplifiers with negative feedback*, SIAM J. Appl. Math., 66 (2005), pp. 468–488.

[7] D. G. Holmes and T. A. Lipo, *Pulse Width Modulation for Power Converters: Principles and Practice*, IEEE Press Ser. Power Engrg., IEEE Press, Piscataway, NJ, 2003.

[8] Z. Song and D. V. Sarwate, *The frequency spectrum of pulse width modulated signals*, Signal Processing, 83 (2003), pp. 2227–2258.

[9] M. Venturini, *A new sine wave in, sine wave out conversion technique eliminates reactive elements*, in Proceedings of the Powercon 7, San Diego, 1980, pp. E3_1–E3_15.

[10] M. Venturini and A. Alesina, *The generalized transformer: A new bidirectional sinusoidal waveform frequency converter with continuously adjustable input power factor*, in Proceedings of the IEEE PESC'80, Atlanta, 1980, pp. 242–252.

[11] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1944.

[12] P. W. Wheeler, J. Rodríguez, J. C. Clare, L. Empringham, and A. Weinstein, *Matrix converters: A technology review*, IEEE Trans. Industr. Electr., 49 (2002), pp. 276–288.

SIAM J. Appl. Math.
Vol. 69, No. 5, pp. 1438–1456

A DOUBLE LAYER SURFACE TRACTION FREE GREEN'S TENSOR[*]

DARKO VOLKOV[†]

Abstract. A double layer Green's tensor for linear elasticity in half space is computed. Traction free conditions on the surface are imposed making this Green's tensor relevant in geophysics for modeling displacements caused by slips on faults. Past attempts at computing related Green's tensors are discussed. Applications to computing displacement fields by integration over fault regions or by use of asymptotic estimates are presented.

Key words. linear elasticity, Green's tensors in half space, traction free surface, cracks in half space, slip along cracks

AMS subject classifications. 74B05, 86-08, 86A17

DOI. 10.1137/080723697

1. Introduction.

Inside a linear elastic region Ω with Lamé coefficients $\lambda > 0$ and $\mu > 0$, a displacement field u satisfies the equation

$$\mu\Delta u + (\lambda+\mu)\nabla \operatorname{div} u = 0 \text{ in } \Omega \tag{1}$$

or alternatively

$$\operatorname{div} \sigma = 0 \text{ in } \Omega, \tag{2}$$

where the stress tensor is given by

$$\sigma_{ij}(u) = \lambda \operatorname{div} u\, \delta_{ij} + \mu(\partial_i u_j + \partial_j u_i).$$

We will use the following notation for stress vectors in the normal direction n throughout this paper:

$$T_n u = \sigma(u)n.$$

The natural basis for $\mathbb{R}^3$ will be denoted by (e_1, e_2, e_3). If Ω is unbounded, a finite energy condition for displacements is required:

$$\int_\Omega \sigma(u) : \epsilon(u) < \infty, \tag{3}$$

where we have used the strain tensor $\epsilon_{ij}(u) = \frac{1}{2}(\partial_i u_j + \partial_j u_i)$ and the dot product between two 3×3 matrices A and B defined by $A : B = \operatorname{tr}(A^T B)$. If Ω is the whole space $\mathbb{R}^3$, it is known since Kelvin that the tensor

$$G_{ij}(x,y) = \frac{1}{8\pi\mu(\lambda+2\mu)}((\lambda+\mu)\partial_{x_i} r \partial_{x_j} r + (\lambda+3\mu)\delta_{ij})\frac{1}{r}, \tag{4}$$

[*] Received by the editors May 8, 2008; accepted for publication (in revised form) October 14, 2008; published electronically February 25, 2009. This work was partially supported by NSF grant DMS 0707421.

http://www.siam.org/journals/siap/69-5/72369.html

[†] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 (darko@wpi.edu).

where $r = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$, satisfies Green's problem

$$\mu \Delta G + (\lambda + \mu) \nabla \operatorname{div} G = -I_3 \delta_y \text{ in } \mathbb{R}^3, \tag{5}$$

where I_3 is the 3×3 identity matrix. In addition G decays at infinity and has finite energy away from the singularity at $x = y$,

$$\int_{\mathbb{R}^3 \setminus B(y,1)} \sigma(G(x,y)) : \epsilon(G(x,y)) dx < \infty. \tag{6}$$

Let Γ be a bounded fault or cut in the space $\mathbb{R}^3$. It is possible to use tensor G to find an integral representation for displacement fields in $\mathbb{R}^3$ that are continuous across Γ and whose stress vector has a given discontinuity (sometimes called jump) across Γ; see [5].

We are interested in this paper in elastic displacement fields in the half space $x_3 < 0$, denoted by $\mathbb{R}^{3-}$, that are traction free on the surface $x_3 = 0$, satisfy some discontinuity condition across a bounded surface Γ in $\mathbb{R}^{3-}$, and decay at infinity while having finite energy. Such displacement fields u can be expressed as integrals on Γ involving Green's tensor M which satisfies

$$\mu \Delta M + (\lambda + \mu) \nabla \operatorname{div} M = -I_3 \delta_y \text{ in } \mathbb{R}^{3-}, \tag{7}$$

$$T_{e_3} M = 0 \text{ on the surface } x_3 = 0, \tag{8}$$

$$M \text{ decays at infinity and } \int_{\mathbb{R}^{3-} \setminus B(y,1)} \sigma(M(x,y)) : \epsilon(M(x,y)) dx < \infty. \tag{9}$$

Mindlin was the first to compute a tensor of this type; see [4]. Sheu performed an analogous computation in the anisotropic case; see [7]. In that same paper he was able to reconstruct displacement fields produced by the 1999 Jiji, Taiwan earthquake using his new Green's tensor.

If u is a finite energy elastic displacement field in the half space $\mathbb{R}^{3-}$ that has zero traction on the surface $x_3 = 0$ and satisfies some discontinuity condition across a bounded surface Γ in $\mathbb{R}^{3-}$, then u can be expressed as the integral over Γ of M against some density which solves an adequate boundary integral equation. These equations on Γ were studied by Martin, Päivärinta, and Rempel in [3].

It might be costly and nontrivial to solve the boundary integral equations discussed in [3]. However, this can be avoided altogether in some cases. Assume that we want to solve for a (finite energy, decaying at infinity) displacement field u such that

$$\mu \Delta u + (\lambda + \mu) \nabla \operatorname{div} u = 0 \text{ in } \mathbb{R}^{3-} \setminus \Gamma, \tag{10}$$

$$T_{e_3} u = 0 \text{ on the surface } x_3 = 0, \tag{11}$$

$$u \text{ is continuous across } \Gamma, \tag{12}$$

$$[T_n u] = f \text{ is a given jump across } \Gamma; \tag{13}$$

then u is given by the integral formula

$$u = \frac{1}{2} \int_\Gamma M f. \tag{14}$$

Note that the free space analogue of problem (10), (12), and (13) is given by the field $u = \frac{1}{2} \int_\Gamma G f$ (see [5]), and from there integral formula (14) is easily conceived in half space. Let us now examine the adjoint problem to (10)–(13), namely, solving for a (finite energy, decaying at infinity) displacement field u such that

$$\mu\Delta u + (\lambda+\mu)\nabla \operatorname{div} u = 0 \text{ in } \mathbb{R}^{3-} \setminus \Gamma, \tag{15}$$

$$T_{e_3} u = 0 \text{ on the surface } x_3 = 0, \tag{16}$$

$$T_n u \text{ is continuous across } \Gamma, \tag{17}$$

$$[u] = g \text{ is a given jump across } \Gamma. \tag{18}$$

We know from [5] that the free space analogue of this problem has the solution $u = \frac{1}{2}\int_\Gamma (T_{n(y)}G)^T g$, where, as previously, G is Kelvin's Green's tensor and $T_{n(y)}$ is the stress vector in the y dependent normal direction $n(y)$. The main result of this paper is to find Green's tensor H such that problem (15)–(18) has the solution $u = \frac{1}{2}\int_\Gamma Hg$ for any smooth tangential vector field on Γ, g. Note that some authors have incorrectly thought that H could be simply given by $(T_{n(y)}M)^T$; this is not true for the operators $T_{e_3(x)}$ and $(T_{n(y)}\cdot)^T$ do not commute when applied to Mindlin's Green's tensor M. We verified this using a symbolic computer software. Actually, this lack of a commutativity property can be understood on simple examples. We discuss three such simple examples in this paper. Steketee was the first author to offer a correct approach on how to compute Green's tensor H; see [6]. Interestingly, he was able to give a full solution only in Fourier space, and he was able to complete his computation in section 7 of his paper [6] in only one particular case.

We now outline the contents of this paper. In section 2 we present our method for computing Green's tensor H. We do not provide any explicit calculations. They are a good order of magnitude more complex than those of Mindlin's because we have to start from the derivative of Kelvin's tensor. It was actually possible for us to perform this calculation thanks to the use of a symbolic calculus software. We indicate the form of the final solution in the appendix. In effect, we are able to provide relatively concise formulas for H only on the surface $x_3 = 0$. For $x_3 < 0$, the simplest form for H is too lengthy to appear in this paper. Note, however, that this formula for H is still malleable using a symbolic calculus software, and it can then be turned into any computer language code. We also demonstrate in section 2 how our Green's tensor relates to two-dimensional (2D) linear elasticity in a half plane. Assuming that displacements occur only in one direction, we are able to recover Green's scalar function in the lower half plane with zero Neumann condition on the surface. In section 3 we explain why under some decay and growth condition, at infinity and near singular points, our Green's tensor is unique, and we use this uniqueness result to verify our (long!) calculation for H. We also include in section 3 a paragraph aimed at understanding why past attempts at computing Green's tensor H were erroneous and why discrepancies were not picked up on numerical data. In section 4 we use our tensor H for the explicit numerical computation of surface displacement fields due to a slip on a crack, or fault, beneath the surface. We note that the exact solution given by $u = \frac{1}{2}\int_\Gamma Hg$ might require a costly computation, which is undesirable in applications where such a direct computation would have to be iterated a large number of times. We found a way to obtain an approximate field $u(x_1, x_2, 0)$ based on asymptotics that just assume that (x_1, x_2) is some distance away from the fault Γ. We also evaluate in section 4 the error incurred in making that approximation. We then discuss in that same section an interesting symmetry property, valid for deep faults.

2. Assembling Green's tensor *H*.

We start from the well-known Kelvin Green's tensor G given by (4). Let $n(y)$ be a y dependent normal direction. We define a double layer potential by setting

$$\tilde{G}(x,y,n) = (T_{n(y)}G(x,y))^T. \tag{19}$$

2.1. The image method. The image method consists of combining $\tilde{G}(x,y,n)$ with terms from

$$(T_{\overline{n}(y)}G(x,\overline{y}))^T,$$

where $\overline{n} = (n_1, n_2, -n_3)$ and $\overline{y} = (y_1, y_2, -y_3)$, in such a way to obtain vanishing traction on the plane $x_3 = 0$, along the x_1 and x_2 directions. More precisely, set

$$\begin{aligned}\tilde{\tilde{G}}_{ij}(x,y) &= \tilde{G}_{ij}(x,y,n) + \tilde{G}_{ij}(x,\overline{y},\overline{n}) \text{ for } 1 \le i \le 3, \quad 1 \le j \le 2,\\ \tilde{\tilde{G}}_{i3}(x,y) &= \tilde{G}_{i3}(x,y,n) - \tilde{G}_{i3}(x,\overline{y},\overline{n}) \text{ for } 1 \le i \le 3.\end{aligned}$$

If g is a smooth vector field on Γ, then $u(x) = \frac{1}{2}\int_\Gamma \tilde{\tilde{G}}(x,y)g(y)$ satisfies (15), (17), and (18), has finite elastic energy, and decays at infinity. However, (16) is only partially satisfied; only the first two components of $T_{e_3}u$ are zero at $x_3 = 0$. Consequently, to find Green's function H, we need to solve three Boussinesq problems with data

$$-F_j := T_{e_3(x)}\tilde{\tilde{G}}_{3j}(x,y)|_{x_3=0}, \quad j = 1,2,3, \tag{20}$$

to compensate for the nonzero $T_{e_3}u \cdot e_3$ term on the surface $x_3 = 0$.

2.2. A Fourier method for solving Boussinesq problems. We find it most efficient to follow the method outlined by Steketee [6]. Recall the definition of Boussinesq half space elasticity problems: find v of finite elastic energy in $\mathbb{R}^{3-}$ such that

$$\mu\Delta v + (\lambda+\mu)\nabla \operatorname{div} v = 0 \text{ in } \mathbb{R}^{3-}, \tag{21}$$

$$T_{e_3}u \cdot e_i = 0 \text{ on the surface } x_3 = 0, \quad i = 1,2, \tag{22}$$

$$T_{e_3}u \cdot e_3 = -F \text{ on the surface } x_3 = 0. \tag{23}$$

The solution to (21)–(23) can be sought in terms of a Galerkin vector $(0,0,\gamma)$, where γ is biharmonic in the lower plane $x_3 < 0$. Indeed, if a displacement field v is in the form

$$v = \left(-\alpha\partial_1\partial_3\gamma,\ -\alpha\partial_2\partial_3\gamma,\ \left[(1-\alpha)\partial_3\partial_3 + \partial_1^2 + \partial_2^2\right]\gamma\right), \text{ where } \alpha = \frac{\lambda+\mu}{\lambda+2\mu}, \tag{24}$$

then it satisfies (21). Note that $\frac{1}{2} < \alpha < 1$. The stress vector $T_{e_3}v$ simplifies as

$$\begin{aligned}\sigma_{13}(v) &= \mu\partial_1(\Delta - 2\alpha\partial_3\partial_3)\gamma,\\ \sigma_{23}(v) &= \mu\partial_2(\Delta - 2\alpha\partial_3\partial_3)\gamma,\\ \sigma_{33}(v) &= \partial_3((\lambda(1-\alpha)+2\mu)\Delta - 2\alpha\mu\partial_3\partial_3)\gamma.\end{aligned}$$

Starting from the problem

$$\Delta^2\gamma = 0 \text{ in } x_3 < 0, \tag{25}$$

$$\sigma_{13}(v) = \sigma_{23}(v) = 0 \text{ on } x_3 = 0, \tag{26}$$

$$\sigma_{33}(v) = -F \text{ on } x_3 = 0, \tag{27}$$

$$\gamma \text{ is bounded as } x_3 \to -\infty, \tag{28}$$

we perform a Fourier transform of γ in the first two variables only. A long calculation leads to

$$\hat{\gamma}(\xi, x_3) = \left(-\frac{1}{2}\,\frac{(-1+2\,\alpha)}{|\xi|^3 \alpha\,\pi^3} + \frac{x_3}{\pi^2 |\xi|^2} \right) \hat{F} \frac{1}{8\,(\mu+\lambda)\,(\alpha-1)} e^{2\,\pi\,|\xi| x_3}, \tag{29}$$

where Fourier transforms are given by

$$\hat{F}(\xi_1, \xi_2) = \iint e^{2\pi i (x_1 \xi_1 + x_2 \xi_2)} F(x_1, x_2) dx_1 dx_2.$$

$\hat{v}$ can be now found according to formula (24). Finally, applying an inverse Fourier transform to $\hat{v}$ will give a finite energy, decaying at infinity, vector field v which satisfies the elasticity equations in $\mathbb{R}^{3-}$ and whose stress vector at the surface $x_3 = 0$ satisfies (26) and (27).

2.3. The Boussinesq solution in our case. Can the vector field v defined in the previous paragraph be given in a closed form? The answer is yes if F is the force coming from adding to Kelvin's tensor its image above the plane $x_3 = 0$ and computing, for each column, the resulting vertical traction at $x_3 = 0$; this will yield Mindlin's tensor. In our case the forcing term F is given by (20); this case involves more terms, of higher degree, compared to those appearing in the derivation of Mindlin's solution. Steketee was able to carry out such a computation in section 7 of his paper [6] in only one particular case. At the time of his work, symbolic algebra software was not available, and this greatly limited investigators' ability to manipulate large expressions. Going back to our work, let us give, for illustration, the expression of F_1, the forcing term for the first Boussinesq problem that we need to solve. F_1 is the ratio of

$$\begin{aligned} \big((2\,\mu+\lambda)\,\mu^2 {x_1}^4 + \left((\mu+2\,\lambda)\,\mu^2 {x_2}^2 - {y_3}^2\,(13\,\lambda+11\,\mu)\,\mu^2 \right) {x_1}^2 + (\lambda-\mu)\,\mu^2 {x_2}^4 \\ + {y_3}^2\,(\mu+2\,\lambda)\,\mu^2 {x_2}^2 + {y_3}^4\,(2\,\mu+\lambda)\,\mu^2 \big) n_1 \\ + \left(3\,\mu^3 x_2 {x_1}^3 + \left(3\,\mu^3 {x_2}^3 - 3\,{y_3}^2\,(4\,\mu+5\,\lambda)\,\mu^2 x_2 \right) x_1 \right) n_2 \\ + \left(-3\,y_3\,(\mu+\lambda)\,\mu^2 {x_1}^3 + \left(-3\,y_3\,(\mu+\lambda)\,\mu^2 {x_2}^2 + 12\,{y_3}^3\,(\mu+\lambda)\,\mu^2 \right) x_1 \right) n_3 \end{aligned}$$

to

$$\left({x_1}^2 + {x_2}^2 + {y_3}^2 \right)^{7/2} \pi\,\mu\,(2\,\mu+\lambda)\,.$$

To compute v given by (24)–(29), for $F = F_j$, $j = 1, 2, 3$, we first computed the Fourier transform $\hat{F}$, which we multiplied by adequate terms to find $\hat{\gamma}$ according to formula (29). We then proceeded to compute the inverse Fourier transform of $\hat{\gamma}$, from which the expression for v, solution to (21)–(23), follows from (24). Corresponding double integrals were evaluated in polar coordinates. A symbolic calculation software had to be used due to the length and complexity of the expressions involved. Of particular importance for polar angle integration was the use of the following integrals:

$$\int_0^{2\pi} e^{iz\cos\theta} \cos p\theta d\theta = (i)^p 2\pi J_p(z), \tag{30}$$

where p is an integer and J_p is the Bessel function of the first kind of order p. This formula can be derived from formula (9.1.21) in [1]. As to integration in radius, a formula for

$$\int_0^\infty \frac{J_q(2\pi\rho r)\rho^p}{(\rho^2 + y_3^2)^{\frac{7}{2}}} d\rho, \quad p = 1, \dots, 5, \quad q = 0, \dots, 4 \quad y_3 < 0, \quad r > 0, \tag{31}$$

was needed. For compuation of inverse Fourier transforms the following was also needed

$$\int_0^\infty e^{2\pi r x_3} J_p(2\pi\rho r) r^q dr \quad p = 0, \dots, 4 \quad q = 0, \dots, 2, \quad x_3 < 0, \quad \rho > 0, \tag{32}$$

Closed forms for (31), (32), albeit intricate, can be computed. The final expression for H is intricate and involves many terms. We discuss it in the appendix.

2.4. Symmetry properties. We first notice that Green's tensor H depends on x_1, y_1, x_2, and y_2 only through $x_1 - y_1$ and $x_2 - y_2$.

2.4.1. Switching the first two coordinates. Let (t_1, t_2, t_3) be a vector in $\mathbb{R}^3$. Denote by $u = (u_1, u_2, u_3)$ the vector $H(t_1, t_2, t_3)$. (u_1, u_2, u_3) is a function of $x_1 - y_1$, $x_2 - y_2$, $x_3 \leq 0$, $y_3 \leq 0$, (n_1, n_2, n_3), (t_1, t_2, t_3), $\lambda > 0$, and $\mu > 0$. The following relations hold:

$$u_1(x_1 - y_1, x_2 - y_2, n_1, n_2, t_1, t_2) = u_2(x_2 - y_2, x_1 - y_1, n_2, n_1, t_2, t_1), \tag{33}$$

$$u_3(x_1 - y_1, x_2 - y_2, n_1, n_2, t_1, t_2) = u_3(x_2 - y_2, x_1 - y_1, n_2, n_1, t_2, t_1). \tag{34}$$

Physically, they express that the first and the second coordinate play the same role for the displacement vector Ht.

2.4.2. Switching the normal vector *n* and the source vector *t*. Computations indicate that the coordinates of Ht depend on the normal vector n and on t only through

$$n_1 t_1, \quad n_2 t_2, \quad n_3 t_3, \quad n_1 t_2 + n_2 t_1, \quad n_1 t_3 + n_3 t_1, \quad n_2 t_3 + n_3 t_2,$$

and, consequently,

$$u(n, t) = u(-n, -t), \tag{35}$$

and

$$u(n, t) = u\left(\frac{t}{|t|}, n|t|\right). \tag{36}$$

Symmetry property (35) corresponds to reversing the orientation on the fault Γ. Symmetry property (36) expresses that the displacements caused by a concentrated slip of vector t, on an infinitesimal fault of normal vector n, are the same as the displacements caused by a concentrated slip of vector $n|t|$, on an infinitesimal fault of normal vector $t/|t|$. We will give in a subsequent section another interpretation of this symmetry property valid for deeper faults of finite size.

2.5. Relation to 2D elasticity. Two dimensional scalar elasticity is the limit model of general elasticity as boundary conditions are constant along a given direction, say, x_2, and displacements take place only in the x_2 direction. We assume here that the fault Γ introduced earlier is linear and infinite in the x_2 direction; thus, a normal vector to Γ satisfies $n_2 = 0$. We then integrate the vector He_2 in x_2 in the range $(-\infty, \infty)$. A long computation leads, after simplification, to the vector

$$\left(0, \frac{(x_1 - y_1)\left({x_3}^2 + {y_3}^2 + (x_1 - y_1)^2\right) n_1 + y_3\left({x_3}^2 - {y_3}^2 - (x_1 - y_1)^2\right) n_3}{\pi\left((x_1 - y_1)^2 + (x_3 - y_3)^2\right)\left((x_1 - y_1)^2 + (x_3 + y_3)^2\right)}, 0\right). \tag{37}$$

Next, we show that this is in agreement with the 2D model. It is known that

$$g(x_1,y_1,x_3,y_3) = -\frac{1}{4}\,\frac{\ln\left((x_1-y_1)^2+(x_3-y_3)^2\right)}{\pi} - \frac{1}{4}\,\frac{\ln\left((x_1-y_1)^2+(x_3+y_3)^2\right)}{\pi}$$

is the half plane $x_3<0$ Green's function for the 2D Laplacian in the x_1, x_3 coordinates that satisfies the zero Neumann condition $\partial_{x_3} g = 0$ at $x_3 = 0$. Note that $\partial_{x_3}\partial_{y_1} g$ and $\partial_{x_3}\partial_{y_3} g$ are also zero at $x_3 = 0$. Computing

$$\partial_{y_1} g\, n_1 + \partial_{y_3} g\, n_3,$$

we find exactly the second coordinate of the vector given in (37).

Remark. The scalar operators ∂_{x_3}, ∂_{y_1}, and ∂_{y_3} do commute. However, we wish to emphasize that, in 3D elasticity, the argument cannot be as simple since the traction operators $T_{e_3(x)}$ and $(T_{e_j(y)}\cdot)^T$ are not commutative.

3. Verification.

We were able to devise a way of verifying our long computation resulting in a closed form for the tensor H.

3.1. A uniqueness theorem.

THEOREM 3.1. *There is a unique tensor $A(x,y)$, for x and y in $\mathbb{R}^{3-}$, whose entries are measurable functions in (x,y) and which satisfies the following equations:*

$$\mu\Delta_x A(x,y) + (\lambda+\mu)\nabla_x \mathrm{div}_x A(x,y) = 0 \text{ in } \mathbb{R}^{3-} \text{ if } x \neq y, \tag{38}$$

$$T_{e_3(x)} A(x,y) = 0 \text{ on the surface } x_3 = 0 \text{ if } y_3 < 0, \tag{39}$$

$$|A(x,y)| \leq \frac{C}{|x|} \text{ as } y \text{ is fixed and } |x| \to \infty, \tag{40}$$

$$|\nabla_x A(x,y)| \leq \frac{C}{|x|^2} \text{ as } y \text{ is fixed and } |x| \to \infty, \tag{41}$$

$$|A(x,y) - \tilde{G}(x,y,n)| \leq C \text{ as } y \text{ is fixed and } x \to y, \tag{42}$$

$$|\nabla_x (A(x,y) - \tilde{G}(x,y,n))| \leq C \text{ as } y \text{ is fixed and } x \to y, \tag{43}$$

where C is a constant independent of x and $\tilde{G}$ is defined by (19).

Proof. It is clear that our Green's tensor $H(x,y)$ satisfies conditions (38)–(43). To show uniqueness, assume that A_1 and A_2 satisfy (38)–(43) and set $\overline{A} = A_1 - A_2$. Then as $\overline{A}(x,y)$ and $\nabla_x \overline{A}(x,y)$ are bounded for x and y, $x \neq y$, in $\mathbb{R}^{3-}$, $\overline{A}$ satisfies the elasticity equations everywhere in $\mathbb{R}^{3-}$. Next, if $\overline{A}_j$ is the jth column of $\overline{A}$, let B_R be the subset of $\mathbb{R}^3$ defined by $\{x : |x| \leq R \text{ and } x_3 \leq 0\}$. Applying conditions (38)–(41) and integrating by parts,

$$\int_{B_R} \varepsilon\left(\overline{A}_j\right) : \sigma\left(\overline{A}_j\right) = \int_{\partial B_R} T_n\left(\overline{A}_j\right)\overline{A}_j.$$

Applying boundary condition (39) and decay at infinity (40)–(41), we find that

$$\int_{\mathbb{R}^{3-}} \varepsilon\left(\overline{A}_j\right) : \sigma\left(\overline{A}_j\right) = 0.$$

$\overline{A}_j$ is then a rigid displacement which, due to the imposed decay at infinity, must be zero. ◻

3.2. Application to verifying our calculation for the tensor $\boldsymbol{H}$. Equations (38) and (39) were verified directly. Conditions (40) and (41) are satisfied with an additional order of magnitude for H, that is,

$$|H(x,y)| \leq \frac{C}{|x|^2} \text{ as } y \text{ is fixed and } |x| \to \infty,$$
$$|\nabla_x H(x,y)| \leq \frac{C}{|x|^3} \text{ as } y \text{ is fixed and } |x| \to \infty.$$

More precisely, each entry of $H(x,y)$ is asymptotically equivalent as $|x| \to \infty$, to the ratio of some homogeneous polynomial of degree 11 in $x_1, x_2, x_3, \sqrt{x_1^2+x_2^2}$, and $\sqrt{x_1^2+x_2^2+x_3^2}$ to $(x_1^2+x_2^2)^3(x_1^2+x_2^2+x_3^2)^{\frac{7}{2}}$.

H also satisfies conditions (42) and (43); they express that $H(x,y)$ and $\tilde{G}(x,y,n)$ have the same type of singularity as y approaches x.

3.3. The problem with past attempts at finding Green's tensor $\boldsymbol{H}$. Some authors have incorrectly thought that H could be simply given by $(T_{n(y)}M)^T$; this is not true for the operators $T_{e_3(x)}$ and $(T_{n(y)}\cdot)^T$ do not commute when applied to Mindlin's Green's tensor M. We verified this fact using a symbolic computer software; note, however, that this lack of a commutativity property can be understood on simple examples. We discuss such simple examples below.

Example 1. For the vector $v(x,y) = (0,0,x_1y_3)$, $T_{e_3(x)}v = (\mu y_3, 0, 0)$. Let A be the 3×3 matrix (v,v,v); that is, each column of A is v. The first column of $(T_{e_3(y)}T_{e_3(x)}A)^T$ is computed to be equal to (μ^2,μ^2,μ^2), while the first column of $T_{e_3(x)}(T_{e_3(y)}A)^T$ is $(0,0,0)$.

Example 2. The following vector field has zero traction derivative in x at the surface $x_3 = 0$:

$$v(x,y) = \left(-\frac{(2\,\mu+\lambda)\,(x_1-y_1)^2}{\lambda}, 0, (x_1-y_1)\,(x_3-y_3) + (x_1-y_1)\,(x_3+y_3)\right).$$

In other words, calculations indicate that $T_{e_3(x)}v = (0,0,0)$ at $x_3 = 0$. Let A be the 3×3 matrix (v,v,v). A computation indicates that $T_{e_3(x)}A$ does not depend on y. Accordingly, $(T_{e_3(y)}T_{e_3(x)}A)^T$ is zero. The first column of $(T_{e_3(y)}A)^T$ is $(-2\mu x_3, -2\mu x_3, -2\mu x_3)$, and, thus, the first column of $T_{e_3(x)}(T_{e_3(y)}A)^T$ is $(-2\mu^2, -2\mu^2, -2\mu(2\mu+\lambda))$.

Example 3. We computed Mindlin's tensor, following a method that proceeds along the same lines as those sketched earlier in this present paper for the calculation of our Green's tensor H: we started from Kelvin's tensor G, which we reflected about the plane $x_3 = 0$, and finally three Boussinesq problems had to be solved. The case of Mindlin's tensor is computationally less intensive, as it involves terms of smaller degree than that of terms involved in the computation of Green's tensor H. Nevertheless, we found it worthwhile to utilize a symbolic computation software for two reasons: first, this certainly reduces chances of obtaining a wrong result; and, second, it makes it more convenient for verifying the final answer. Verification was made by validating elasticity equations, and decay at infinity, and order of growth near singularities and checking traction free boundary conditions.

In the end we found that although $T_{e_3(x)}M(x,y)$ is zero at $x_3 = 0$, $T_{e_3(x)}$ $(T_{e_3(y)}M(x,y))^T$ is not zero at $x_3 = 0$, whereas previously $M(x,y)$ was Mindlin's tensor. The reason why this was not picked up in previous studies in natural sciences may lie in the following observation: $M(x,y)$ is homogeneous in (x,y) of degree -1, and, consequently, $T_{e3(x)}(T_{e3(y)}M(x,y))^T$ is homogeneous in (x,y) of degree -3. If this tensor is evaluated for y_3 at some depth d beneath the surface $x_3=0$, the error in surface traction $T_{e3(x)}(T_{e3(y)}M(x,y))^T$ decays as d^{-3}, which is one order of magnitude smaller than displacements on the surface. In other words, using $T_{e3(y)}M(x,y)$ for computing surface displacements due to a slip on a fault may lead to gross discrepancies, only if that fault is shallow and for those surface points close enough to the fault.

4. Application: Computations of displacement fields caused by a slip along a fault. In this section we use Green's tensor H introduced in this paper to compute displacement fields u due to a slip on a fault Γ, in the half space $x_3<0$, with traction free conditions on the surface $x_3=0$. In other words, u satisfies (15)–(18). This equation for u plays an important role in geophysics models. It may be used in the study of quasi-static displacements near a fault during a "silent earthquake" episode. Accounts of silent earthquakes in subduction zones near Japan [12] and New Zealand, Alaska, and Mexico [13, 11] were recently reported in the literature. This equation for u may also be used to study the nucleation phase (occurring after destabilization of faults and before the onset of seismic waves) for dynamically active faults. The earthquake nucleation phase, which precedes dynamic rupture, was uncovered by detailed seismological observations [14, 16] and identified in laboratory experiments [15, 17].

Typical length scales attached to faults observed in nature range from 0 to 100 kilometers for depth and 1 to 100 kilometers for length. During destabilization, slip on faults are on the order of 1 to 100 meters. Accordingly, in all numerical simulations in this section, we choose one kilometer to be the unit length for spatial coordinates, while surface displacements are given in meters. The Lamé coefficients are set to be $\lambda=\mu=1$, a common choice in geophysics. This choice of λ and μ is not necessary for our computations to run faster or more accurately. It was rather made in order to facilitate comparisons to other pieces of work.

4.1. Exact field in two numerical examples. Recall that the displacement field u due to a slip g on a fault Γ, in the half space $x_3<0$, with traction free conditions on the surface $x_3=0$, that is, the solution to problem (15)–(18), can be expressed as $u=\frac{1}{2}\int_\Gamma Hg$. We compute in this section the displacement u on the surface $x_3=0$ using this integral formula in two examples.

These two examples involve the same fault geometry: Γ is contained in the plane normal to the vector $(1,0,1)$ and is bounded by an ellipse centered at $(0,0,-2)$. In local coordinates the ellipse has the equation $\tilde{y_1}^2+(\tilde{y_2}/5)^2=1$, where local coordinates are related to the original coordinates by

$$(44)\qquad y=\begin{pmatrix} s & 0 & s \\ 0 & 1 & 0 \\ -s & 0 & s \end{pmatrix}\tilde{y}+\begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix}, \quad s=\frac{1}{\sqrt{2}}.$$

A sketch of this cross section appears in Figure 1.

In the first example slip occurs only in the e_2 direction. In local coordinates the slip was picked to be $g=C_1\sqrt{1-\tilde{y_1}^2-(\tilde{y_2}/5)^2}e_2$, where the constant C_1 was adjusted in such a way that the total slip $\int_\Gamma g$ be of norm 1; see Figure 2 for a plot of resulting surface displacements. We wish to emphasize that this choice of slip g is not arbitrary; it corresponds to the expected dominant profile of slip occurring

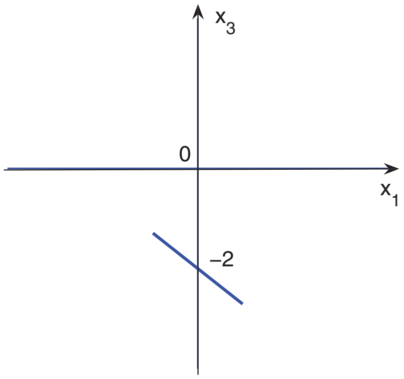


FIG. 1. *A cross section of the fault Γ involved in the first two numerical examples. The cross section is in the plane $x_2 = 0$. The small axis of the ellipse appears as the line segment in the cross section plane.*

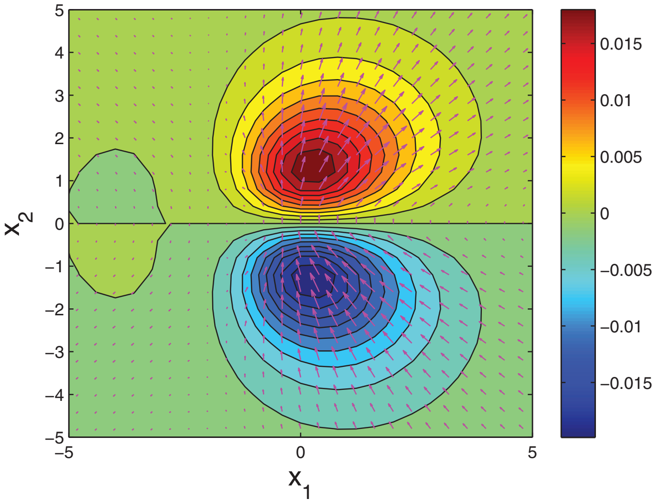


FIG. 2. *Computed surface displacement for the elliptic geometry and slip in the e_2 direction considered in the first example. In local coordinates the slip was picked to be $g = C_1\sqrt{1 - \tilde{y_1}^2 - (\tilde{y_2}/5)^2}e_2$, where the constant C_1 was adjusted in such a way that the total slip $\int_\Gamma g$ be of norm 1. The e_1 and e_2 components of $u(x_1, x_2)$ are represented as a planar vector field using arrows, while the e_3 component is sketched on the same graph using a color contour map.*

during the destabilization process of a fault. A complete theory for that process has been studied in the 2D case; see [2], [8], [9]. In the 3D case a complete theory is still being investigated, but analysis of relevant hypersingular operators suggests that slip occurring during the destabilization process of a fault decays toward the edge of the fault as the square root of the distance to this edge.

In the second example the slip does not have constant direction and is picked to be, in local coordinates, $g = C_2(-2m^{3/2}, m^{1/2}, 2m^{3/2})$, where $m = 1 - \tilde{y_1}^2 - (\tilde{y_2}/5)^2$ and the constant C_2 was adjusted, as previously, in such a way that the total slip $\int_\Gamma g$ still be of norm 1; see Figure 3 for a plot of resulting surface displacements.

4.2. Approximate field. Solving (15)–(18) by integrating $u = \frac{1}{2}\int_\Gamma Hg$ might be costly in number of operations, which is undesirable in applications where such a

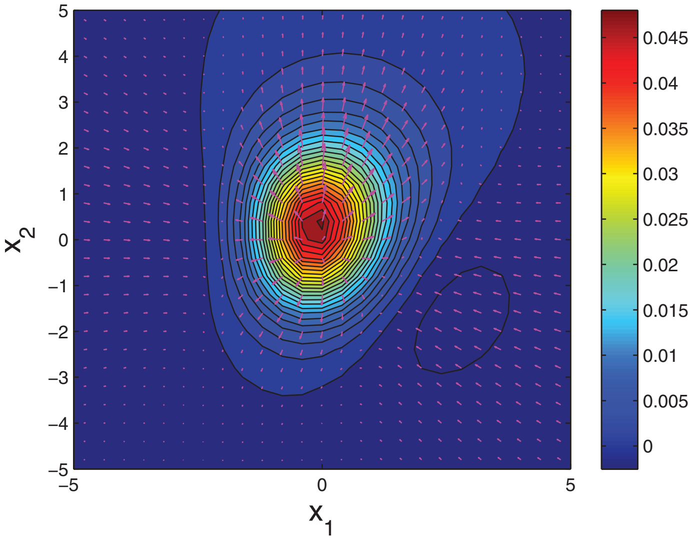


FIG. 3. *Computed surface displacements in the second example. The geometry of the fault is the same as in the first example; however, the slip does not in this case have constant direction and is picked to be, in local coordinates,* $g = C_2(-2m^{3/2}, m^{1/2}, 2m^{3/2})$*, where* $m = 1 - (\tilde{y_1}/5)^2 - \tilde{y_2}^2$ *and the constant* C_2 *was adjusted in such a way that the total slip* $\int_\Gamma g$ *be of norm* 1.

direct computation would have to be iterated a large number of times. We found a way to obtain a reasonable approximation to the field $u(x_1, x_2, 0)$ based on asymptotics that just assume that (x_1, x_2) is some distance away from the fault Γ. Suppose that the fault Γ is centered at the point (a, b, c) where $c < 0$. To obtain a simpler formula for the surface displacement $u(x_1, x_2, 0)$ we now assume that either the surface point (x_1, x_2) is far enough from (a, b) or $|c|$ is large enough. Thus, we may write

$$\begin{aligned} H(x_1, x_2, 0, y_1, y_2, y_3) &= H(x_1 - y_1, x_2 - y_2, 0, 0, 0, y_3) \\ &= H(x_1 - a, x_2 - b, 0, 0, 0, c) + O\left(\frac{1}{(x_1^2 + x_2^2 + c^2)^{3/2}}\right) \end{aligned}$$

as long as (y_1, y_2, y_3) remains on the fault Γ. From there, integrating over Γ,

$$u(x_1, x_2, 0) \simeq H(x_1 - a, x_2 - b, 0, 0, 0, c)\frac{1}{2}\int_\Gamma g(y_1, y_2, y_3)dy. \tag{45}$$

Setting $(t_1, t_2, t_3) = \frac{1}{2}\int_\Gamma g(y_1, y_2, y_3)dy$, we obtain

$$u(x_1, x_2, 0) \simeq H(x_1 - a, x_2 - b, 0, 0, 0, c)(t_1, t_2, t_3). \tag{46}$$

The vector $t := (t_1, t_2, t_3)$ can be interpreted as half the average slip on Γ times the area of Γ. We will call $2t$ the total slip on Γ.

We now proceed to demonstrate numerically the accuracy of approximation (46). We plot the relative L^2 error incurred in making the approximation (46) against depth for three different geometries in Figure 4. The L^2 error was computed on the surface $x_3 = 0$ in a square $[-10, 10] \times [-10, 10]$. In each case the fault was contained in the plane normal to the vector $(1, 0, 1)$ and passing through the center $(0, 0, c)$, where $|c|$ is the depth. Depth ranged from 2 to 20 in these numerical runs. Slip

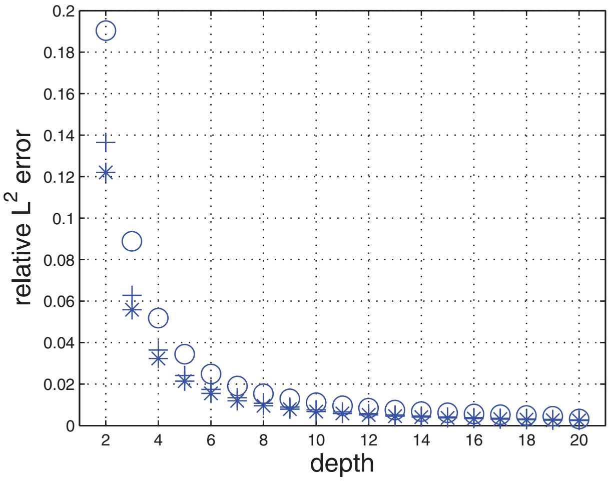


FIG. 4. *The relative L^2 error incurred by making the approximation* (46) *plotted against depth for the three different fault geometries discussed in section* 4.2.

was set to occur in the e_2 direction. Total slip was computed in order to apply formula (46). The plus markers correspond to a square geometry for Γ with edges of length 2. In local coordinates $(\tilde{y_1}, \tilde{y_2})$ centered on the fault, the slip was picked to be $\sqrt{(1-|\tilde{y_1}|)(1-|\tilde{y_2}|)}$. Local coordinates are now related to the original coordinates by

$$y = \begin{pmatrix} s & 0 & s \\ 0 & 1 & 0 \\ -s & 0 & s \end{pmatrix} \tilde{y} + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}, \quad s = \frac{1}{\sqrt{2}}.$$

The star markers correspond to Γ being a circle of radius 1. In local coordinates the slip was picked to be $\sqrt{1-\tilde{y_1}^2-\tilde{y_2}^2}$. The circular markers correspond to an elliptic geometry for Γ. The equation of the ellipse was picked to be, in local coordinates, $\tilde{y_1}^2+(\tilde{y_2}/5)^2=1$. In local coordinates the slip was picked to be $\sqrt{1-\tilde{y_1}^2-(\tilde{y_2}/5)^2}$. The largest error is found for the most shallow faults, that is, for $|c|=2$, and ranges from 12 to 19%, depending on geometry. We sketched the exact and approximated fields for faults at depth 2 in the elliptic geometry case in Figures 2 and 5. It appears that even at that shallow depth the exact and approximated profiles exhibit very similar patterns.

4.3. A symmetry property for deeper faults.

Denote by $\tilde{u}$ the approximate surface displacement obtained by application of asymptotic formula (46). Accordingly, $\tilde{u} = Ht$. The coordinates of Ht are given in the appendix. Recall that symmetry property (35) corresponds to reversing the orientation on the fault Γ. A consequence of symmetry property (36) is that, for deeper faults, a total slip t on a planar fault Γ of normal vector n produces approximately the same surface displacements as in the "reversed case" of a total slip $n|t|$ on a planar fault Γ of normal vector $t/|t|$. Of course, this equivalence does not hold for shallow faults at surface points close to the fault.

Due to the expression for $H(x_1-a, x_2-b, 0, 0, 0, c)$ it turns out that $u(x_1, x_2)$ is

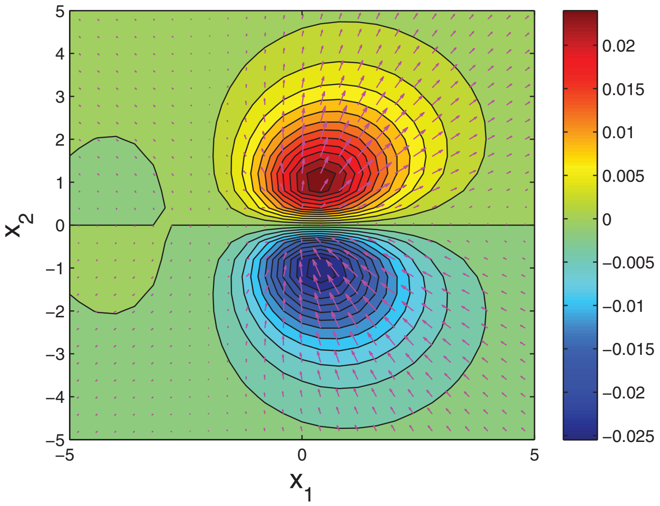


FIG. 5. *Approximate surface displacement for the elliptic geometry from Figure* 4 *at depth* 2. *The approximation was obtained by applying formula* (46). *The computed exact field is sketched in Figure* 2.

a function that depends on n and t only through s_0, s_1, s_2, s_3, and s_4 defined by

$$s_0 := n_1t_1 + n_2t_2, \tag{47}$$

$$s_1 := n_2t_3 + n_3t_2, \tag{48}$$

$$s_2 := n_1t_3 + n_3t_1, \tag{49}$$

$$s_3 := n_1t_2 + n_2t_1, \tag{50}$$

$$s_4 := n_1t_1 - n_2t_2. \tag{51}$$

Thus, one might wonder whether additional symmetries akin to (36) hold. The following proposition explains in detail how s_0, s_1, s_2, s_3, and s_4 relate to n and t if n and t are perpendicular. Note that, in the destabilization process of faults, previous studies have shown that the slip is tangential to the fault, so n and t are indeed perpendicular in that case.

PROPOSITION 4.1. *Assume that* $n = (n_1, n_2, n_3)$ *and* $t = (t_1, t_2, t_3)$ *are two orthogonal vectors in space such that* $|n| = 1$ *and* $|t| \neq 0$. *Given* s_0, s_1, s_2, s_3, *and* s_4 *defined by* (47)–(51) *exactly four different pairs* (n, t) *can be reconstructed. If* $(\tilde{n}, \tilde{t})$ *is one reconstructed pair, the other three are* $(-\tilde{n}, -\tilde{t})$, $(\frac{\tilde{t}}{|\tilde{t}|}, \tilde{n}|\tilde{t}|)$, *and* $(-\frac{\tilde{t}}{|\tilde{t}|}, -\tilde{n}|\tilde{t}|)$.

Proof. Form the matrix

$$D = \begin{pmatrix} 2n_1t_1 & s_3 & s_2 \\ s_3 & 2n_2t_2 & s_1 \\ s_2 & s_1 & 2n_3t_3 \end{pmatrix},$$

and notice that $D = nt^T + tn^T$. Denote $t' = t/|t|$ and $D' = D/|t| = nt'^T + t'n^T$. Since n and t' are orthogonal of norm 1, we have that $D't' = n$ and $D'n = t'$ from where it follows that $D'(n - t') = -(n - t')$ and $D'(n + t') = n + t'$. We conclude that $n - t'$ and $n + t'$ are eigenvectors for D for the respective eigenvalues $-|t|$ and $|t|$. Notice also that $n \times t$ is an eigenvector for the eigenvalue 0.

To reconstruct n and t, as D is real symmetric and has zero trace and determinant, we may denote by $-\alpha, 0, \alpha$ (with $\alpha > 0$) the eigenvalues of D.

To find a pair (n, t) from the symmetric matrix D we have to find two vectors v_1 and v_2 of norm $\sqrt{2}$ such that $Dv_1 = -\alpha v_1$ and $Dv_2 = \alpha v_2$. As D is symmetric v_1 and v_2 are orthogonal. If we then set $n = \frac{v_1 - v_2}{2}, t' = \frac{v_1 + v_2}{2}$, and $t = \alpha t'$, it is clear that the coordinates of (n, t) will satisfy (47)–(51) and so will the coordinates of the other pairs $(-n, -t), (\frac{t}{|t|}, n|t|)$, and $(-\frac{t}{|t|}, -n|t|)$.

Finally, we show that these are the only four solutions. This is because a basis for the eigenspace attached to the eigenvalue $-\alpha$ containing vectors of length $\sqrt{2}$ can only be given by either v_1 or $-v_1$, and a basis for the eigenspace attached to the eigenvalue α containing vectors of length $\sqrt{2}$ can only be given by either v_2 or $-v_2$; this gives a total number of four combinations. □

Remark. Assume that s_0, s_1, s_2, s_3, and s_4 are any five real numbers. Can we find two orthogonal vectors n and t such that $|n| = 1$ and (47)–(51) are satisfied? Forming the matrix

$$E = \begin{pmatrix} s_0 + s_4 & s_3 & s_2 \\ s_3 & s_0 - s_4 & s_1 \\ s_2 & s_1 & -2s_0 \end{pmatrix}$$

this is possible if and only if $\det(E) = 0$. Indeed, the condition $\det(E) = 0$ is necessary since $E(n \times t) = 0$. Conversely, as E is real and symmetric and its trace is zero, if $\det(E) = 0$, the eigenvalues of E must be $-\alpha, 0, \alpha$ for some $\alpha > 0$ unless $s_0 = s_1 = s_2 = s_3 = s_4 = 0$, in which case $E = 0$. The assertion then follows from the previous proposition.

To illustrate numerically the symmetry property, we computed the surface displacements $u(x_1, x_2)$ arising from the slip on the fault Γ of equation $\tilde{y_1}^2 + (\tilde{y_2}/5)^2 = 1$, in local coordinates, at depth $c = -2$ and normal to the vector $(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$, where new and original coordinates are again related by (44). We imposed on that fault the slip $C_1(-2m^{3/2}, m^{1/2}, 2m^{3/2})$, where $m = 1 - \tilde{y_1}^2 - (\tilde{y_2}/5)^2$, in local coordinates, and the constant C_1 was computed such that the norm of the total slip $2t$ was 1. The computed profile appears in Figure 3.

Next, we proceeded to find the surface displacements for a reversed geometry, that is, the surface displacements $u'(x_1, x_2)$ arising from the slip on the fault Γ of equation $\tilde{y_1}^2 + (\tilde{y_2}/5)^2 = 1$, in local coordinates, at depth $c = -2$ and normal to the vector t with imposed slip $C_2((m/2)^{1/2} + 2m^{3/2})(1, 0, 1)$, where $m = 1 - \tilde{y_1}^2 - (\tilde{y_2}/5)^2$, in local coordinates, and the constant C_2 was computed such that the norm of the total slip was 1. The computed profile for $u'(x_1, x_2)$ appears in Figure 6.

We also computed the surface displacements $u''(x_1, x_2)$ obtained by application of asymptotic formula (46). The computed profile for $u''(x_1, x_2)$ appears in Figure 7.

Finally, we compared relative differences in the L^2 norm for u, u', and u'' where the surface domain of integration is $[-10, 10] \times [-10, 10]$. For example, the relative difference of u to u' is $\sqrt{\int |u - u'|^2 / \int |u|^2}$, where all integrals are over $[-10, 10] \times [-10, 10]$. We computed these relative differences for two depths $c = -2$ and $c = -20$, and we placed them in Table 1.

5. Conclusion. We have computed in this paper a double layer Green's tensor for linear elasticity in half space, with traction free conditions on the surface. Our approach starts from Kelvin's free space tensor: we first took a traction derivative, and then we made a long calculation whose goal was to derive additional terms accounting for the traction free boundary condition. We indicated the form of the final solution in the appendix. In effect, we are able to provide relatively concise formulas for H

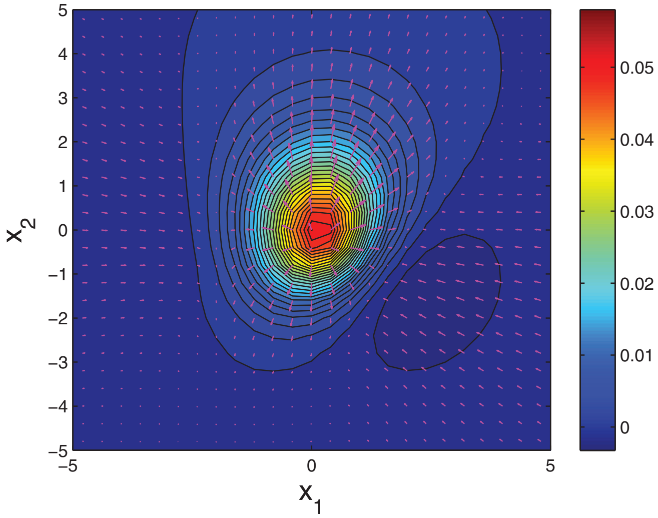


FIG. 6. *Surface displacements $u'(x_1, x_2)$ arising from the slip on the fault* Γ *of equation* $\tilde{y_1}^2 + (\tilde{y_2}/5)^2 = 1$*, in local coordinates, at depth* $c = -2$ *and normal to the vector* t *with imposed slip* $C_2((m/2)^{1/2} + 2m^{3/2})(1, 0, 1)$*, where* $m = 1 - (\tilde{y_1}/5)^2 - \tilde{y_2}^2$*, in local coordinates, and the constant* C_2 *is a computed constant ensuring that the norm of the total slip be* 1*. We observe that the displacement pattern is similar to the one plotted in Figure* 3*.*

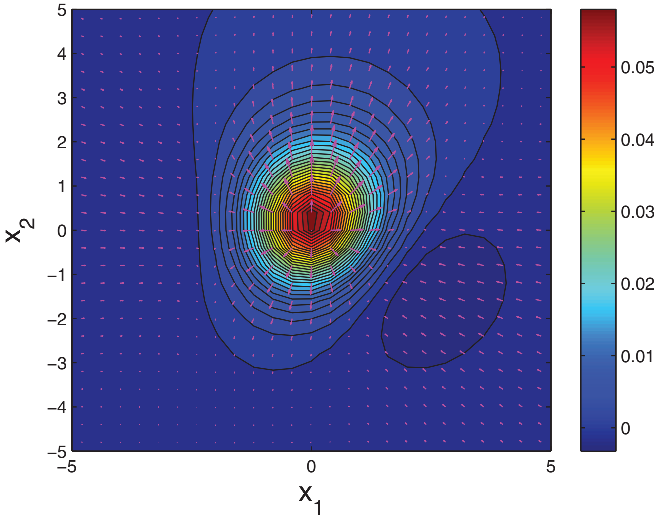


FIG. 7. *Surface displacements $u''(x_1, x_2)$ obtained by application of asymptotic formula* (46) *to either the case relative to Figure* 3 *or to the case relative to Figure* 6*.*

only on the surface $x_3 = 0$. For $x_3 < 0$, the simplest form for H is too lengthy, but still manageable on a computer system.

We also showed that simply starting from Mindlin's half space tensor and then taking a traction derivative leads to an incorrect result: the traction free condition on the surface is lost. This is due to the fact that traction operators do not commute. We illustrated this lack of commutativity on simple examples.

We demonstrated how our Green's tensor relates to 2D linear elasticity in a half

Table 1

Relative differences in the L^2 norm for u, u', and u'' where the domain of integration is $[-10,10] \times [-10,10]$. For example, the relative difference of u to u' is $\sqrt{\int |u-u'|^2 / \int |u|^2}$, where all integrals are over $[-10,10] \times [-10,10]$. We computed these relative differences for two depths $c=-2$ and $c=-20$.

	u to u'	u to u''	u' to u''
$c=-2$	.1814	.1741	.1738
$c=-20$	.002265	.002103	.001984

plane. That case reduces to recovering Green's scalar function in the lower half plane with zero Neumann condition on the surface. We also explained why under some decay and growth condition our Green's tensor is unique, and we use this uniqueness result to verify our calculation for H. Finally, we used our tensor H for the explicit numerical computation of surface displacement fields due to a slip on a crack, or fault, beneath the surface. As the exact solution might require an intensive computation, we found a way to obtain an approximate field $u(x_1, x_2, 0)$ based on asymptotics assuming only that (x_1, x_2) is some distance away from the fault Γ. This led to a discussion on an interesting symmetry property, valid for deeper faults.

The half space setting considered in this paper plays an important role in geophysics, where the traction free plane at the boundary models the surface of the Earth. This geometry may also be helpful in material science at adequate length scales. We have shown in this paper the expression in closed form for H on the surface and its use for efficiently approximating surface fields due to a slip on the fault. In another paper we will demonstrate how one can take advantage of those approximate closed form expressions for surface displacements, in order to solve the fault inverse problem: given a surface displacement field u, can one recover the fault and the slip that gave rise to u? We provide a positive answer to a regularized version of that problem. Our recovery method combines algebraic manipulations on the approximate closed form expressions for surface displacements to minimization techniques; see [10].

6. Appendix.

Instead of giving formulas for each entry of the matrix H, it is advantageous to write out formulas for the coordinates $H\,t$, where t is the vector (t_1, t_2, t_3). We only present in this appendix formulas at $x_3 = 0$; the idea is to give a feel for the different terms involved. The complete formula for $x_3 < 0$ is best left within a computer code.

It proves convenient to introduce polar surface coordinates.

6.1. If $(x_1 - y_1)^2 + (x_2 - y_2)^2 = 0$.

The three coordinates of $H\,t$ are then, respectively,

$$0,$$

$$0,$$

$$\frac{-\mu(n_1 t_1 + n_2 t_2) + 6(\lambda+\mu) n_3 t_3}{4\pi {y_3}^2 (\lambda+\mu)}.$$

6.2. If $(x_1 - y_1)^2 + (x_2 - y_2)^2 > 0$.

We set

$$\rho = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2},$$
$$d = \sqrt{\rho^2 + y_3^2},$$

$$\begin{aligned} c &= \frac{x_1 - y_1}{\rho}, \\ s &= \frac{x_2 - y_2}{\rho}. \end{aligned}$$

The first coordinate of $H\,t$ is the ratio of

$$\alpha n_1 t_1 + \beta n_2 t_2 + \gamma n_3 t_3 + \delta(n_1 t_2 + n_2 t_1) + \epsilon(n_1 t_3 + n_3 t_1) + \zeta(n_2 t_3 + n_3 t_2) \tag{52}$$

to

$$(\lambda + \mu)\pi\, \rho^3 d^5, \tag{53}$$

where, setting

$$A = \left(y_3 d \rho^4 + \left(\frac{5}{2}\, {y_3}^4 + 2\, {y_3}^3 d \right) \rho^2 + {y_3}^6 + {y_3}^5 d \right) \mu\,,$$

$\alpha, \beta, \gamma, \delta, \gamma, \epsilon$, and ζ are given by

$$\begin{aligned} \frac{\alpha}{c} &= -A\left(4\,c^2 - 3\right) + \left(\frac{3}{2}\,\lambda\, c^2 + \mu \right) \rho^6 - \frac{1}{2}\,\mu\, \left(15\,c^2 - 11\right) {y_3}^2 \rho^4, \\ \frac{\beta}{c} &= A\left(4\,c^2 - 3\right) - \frac{3}{2}\,\lambda\, \left(c^2 - 1\right) \rho^6 + \frac{3}{2}\,\mu\, \left(5\,c^2 - 4\right) {y_3}^2 \rho^4, \\ \frac{\gamma}{c} &= \frac{3}{2}\,(\lambda + \mu) {y_3}^2 \rho^4, \\ \frac{\delta}{s} &= -A\left(4\,c^2 - 1\right) + \left(\frac{3}{2}\,\lambda\, c^2 + \frac{1}{2}\,\mu \right) \rho^6 - \frac{1}{2}\,\mu\, \left(15\,c^2 - 4\right) {y_3}^2 \rho^4, \\ \epsilon &= -\frac{3}{2}\,\rho^5 y_3 c^2 (\lambda + \mu), \\ \zeta &= -\frac{3}{2}\,(\lambda + \mu) c s y_3 \rho^5. \end{aligned}$$

The second coordinate of $H\,t$ is also in the form of a ratio of (52) to (53), where this time $\alpha, \beta, \gamma, \delta, \gamma, \epsilon$, and ζ are given by

$$\begin{aligned} \frac{\alpha}{s} &= -A\left(4\,c^2 - 1\right) + \frac{3}{2}\,\lambda\, c^2 \rho^6 - \frac{3}{2}\, \left(5\,c^2 - 1\right) \mu\, {y_3}^2 \rho^4, \\ \frac{\beta}{s} &= A\left(4\,c^2 - 1\right) + \left(\frac{3}{2}\,\lambda + \mu - \frac{3}{2}\,\lambda\, c^2 \right) \rho^6 + \frac{1}{2}\,\mu\, \left(-4 + 15\,c^2\right) {y_3}^2 \rho^4, \\ \frac{\gamma}{s} &= \frac{3}{2}\,(\lambda + \mu) {y_3}^2 \rho^4, \\ \frac{\delta}{c} &= A\left(4\,c^2 - 3\right) + \left(\frac{3}{2}\,\lambda + \frac{1}{2}\,\mu - \frac{3}{2}\,\lambda\, c^2 \right) \rho^6 + \frac{1}{2}\,\mu\, \left(15\,c^2 - 11\right) {y_3}^2 \rho^4, \\ \epsilon &= -\frac{3}{2}\,(\lambda + \mu) c s y_3 \rho^5, \\ \zeta &= \frac{3}{2}\,\rho^5 y_3 \left(c^2 - 1\right) (\lambda + \mu). \end{aligned}$$

Setting

$$B = \left(\frac{1}{2}\, d\rho^5 + dy_3{}^2\rho^3 + \left(\frac{1}{2}\, y_3{}^5 + \frac{1}{2}\, y_3{}^4 d\right)\rho\right)\mu\,,$$

the third coordinate of $H\,t$ is also in the form of a ratio of (52) to (53), where this time $\alpha, \beta, \gamma, \delta, \gamma, \epsilon$, and ζ are given by

$$\alpha = -B\left(-1 + 2\,c^2\right) + \left(-3\,\mu\,c^2 - \frac{3}{2}\,\lambda\,c^2 + \mu\right) y_3\rho^5 - \frac{1}{2}\,\mu\,\left(5\,c^2 - 3\right) y_3{}^3\rho^3,$$
$$\beta = B\left(-1 + 2\,c^2\right) + \left(-2\,\mu + \frac{3}{2}\,\lambda\,c^2 - \frac{3}{2}\,\lambda + 3\,\mu\,c^2\right) y_3\rho^5 + \frac{1}{2}\,\mu\,\left(-2 + 5\,c^2\right) y_3{}^3\rho^3,$$
$$\gamma = -\frac{3}{2}\,(\lambda + \mu) y_3{}^3\rho^3,$$
$$\frac{\delta}{sc} = \left(\left(-\frac{3}{2}\,\lambda - 3\,\mu\right) y_3 - d\mu\right)\rho^5 + \left(-\frac{5}{2}\,\mu\,y_3{}^3 - 2\,d\mu\,y_3{}^2\right)\rho^3 + \left(-d\mu\,y_3{}^4 - \mu\,y_3{}^5\right)\rho,$$
$$\frac{\epsilon}{c} = \frac{3}{2}\,(\lambda + \mu) y_3{}^2\rho^4,$$
$$\frac{\zeta}{s} = \frac{3}{2}\,(\lambda + \mu) y_3{}^2\rho^4.$$

Remark. Denote by (u_1, u_2, u_3) the coordinates of Ht, whose expressions were given above. From (33) and (34) the following symmetry properties must hold:

$$u_1(s, c, n_1, n_2, t_1, t_2) = u_2(c, s, n_2, n_1, t_2, t_1),$$
$$u_3(s, c, n_1, n_2, t_1, t_2) = u_3(c, s, n_2, n_1, t_2, t_1).$$

These symmetry properties can be directly verified from the previous formulas using that $s^2 + c^2 = 1$.

Acknowledgments. The author wishes to thank Prof. I. R. Ionescu for his offering of important suggestions and helpful comments.

REFERENCES

[1] M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* Dover, New York, 1992.

[2] C. Dascalu and I. R. Ionescu, *Slip weakening friction instabilities: Eigenvalue analysis,* Math. Models Methods Appl. Sci., 14 (2004), pp. 439–459.

[3] P. A. Martin, L. Päivärinta, and S. Rempel, *A normal crack in an elastic half-space with stress-free surface. English summary,* Math. Methods Appl. Sci., 16 (1993), pp. 563–579.

[4] R. D. Mindlin, *Force at a point in the interior of a semi infinite solid,* Phys., 7 (1936), pp. 195–202.

[5] V. Z. Parton and P. I. Perlin, *Integral Equations in Elasticity,* Mir, Moscow, 1982. MR 509209 (80a:73001)

[6] J. A. Steketee, *On Volterra's dislocations in a semi infinite elastic medium,* Canad. J. Phys., 36 (1958), pp. 192–205.

[7] G. Y. Sheu, *Deformations caused by the movements of shear and tensile faults,* Int. J. Numer. Anal. Methods Geomech., 25 (2001), pp. 1175–1193.

[8] I. R. Ionescu and D. Volkov, *An inverse problem for the recovery of active faults from surface observations,* Inverse Problems, 22 (2006), pp. 2103–2121.

[9] I. R. Ionescu and D. Volkov, *Earth surface effects on active faults: An eigenvalue asymptotic analysis*, J. Comput. Appl. Math., 220 (2008), pp. 143–162.

[10] I. R. Ionescu and D. Volkov, *Detecting tangential dislocations on planar faults from traction free surface observations*, Inverse Problems, 25 (2009), 015012.

[11] V. Kostoglodov, S. K. Singh, J. A. Santiago, S. I. Franco, K. M. Larson, A. R. Lowry, and R. Bilham, *A large silent earthquake in the Guerrero seismic gap, Mexico,* Geophys. Res. Lett., 30 (2003) citation 1807.

[12] T. Sagiya and S. Ozawa, *Anomalous transient and silent earthquakes along the Nankai Trough subduction zones,* Seismol. Res. Lett., 73 (2002), pp. 234–235.

[13] A. R. Lowry, K. M. Larson, V. Kostoglodov, and O. Sanchez, *The fault slip budget in Guerrero, Southern Mexico,* Geophys. J. Int., 200 (2005), pp. 1–15.

[14] Y. Iio, *Slow initial phase of the P-wave velocity pulse generated by microearthquakes,* Geophys. Res. Lett., 19 (1992), pp. 477–480.

[15] J. H. Dieterich, *A model for the nucleation of earthquake slip*, in Earthquake Source Mechanics, Geophys. Monogr. Ser. 37, S. Das, J. Boatwright, and C.H. Scholz, eds., American Geophysical Union, Washington, D. C., 1986, pp. 37–47.

[16] W. L. Elsworth and G. C. Beroza, *Seismic evidence for an earthquake nucleation phase,* Sci., 268 (1995), pp. 851–855.

[17] M. Ohnaka, Y. Kuwakara, and K. Yamamoto, *Constitutive relations between dynamic physical parameters near a tip of the propagation slip during stick-slip shear failure,* Tectonophysics, 144 (1987), pp. 109–125.

SIAM J. Appl. Math.
Vol. 69, No. 5, pp. 1457–1476

A TWO-DIMENSIONAL DIFFUSION APPROXIMATION FOR A LOSS MODEL WITH TRUNK RESERVATION*

CHARLES KNESSL[†] AND JOHN A. MORRISON[‡]

Abstract. We consider a loss model with C servers, and arriving customers are split into two classes. Of the C servers, R may be used only by the high priority class. Thus if a high priority customer sees all C servers occupied, then that customer is lost, while a low priority customer is lost if $\geq C - R$ servers are occupied. Assuming Poisson arrivals of both customer types and exponential service, we study the problem asymptotically, with $C \to \infty$ and the arrival rates comparably large. We assume that the total load is roughly equal to the number of servers, and we obtain a two-dimensional diffusion equation satisfied by the joint steady state probability distribution of the numbers of servers occupied by the two customer classes. We analyze this equation by a combination of analytic and numerical methods. Our singular perturbation analysis makes certain assumptions about not only the forms of various asymptotic expansions but also the asymptotic matching between different scales.

Key words. asymptotics, diffusion, trunk reservation

AMS subject classifications. 60K30, 90B22, 98M20

DOI. 10.1137/070698166

1. Introduction.

We consider the following queueing model. There are C servers and two customer types. The high (resp., low) priority customers arrive according to a Poisson process of rate λ (resp., ν). The service times for the high and low priority customers are all exponentially distributed, with respective means 1 and $1/\kappa$. Of the C servers, R are reserved for the high priority customers. If a high priority customer arrives to see all C servers occupied, that customer is lost (or blocked). If a low priority customer arrives to see at least $C - R$ servers occupied, then that customer is lost.

This model is referred to as "trunk reservation." The concept of trunk reservation is of fundamental importance in circuit-switched networks. On any link of the network, which has a fixed number of circuits, some of the circuits may be reserved for the primary traffic, which is offered directly to the link. Secondary traffic, which is rerouted because of a busy link on its direct route, is accepted on an alternate link only if there is an unreserved link available.

In our notation, the servers are circuits, high priority customers are primary calls, and low priority customers are secondary calls. A circuit is held for the duration of the call. An arriving call is blocked and lost if a circuit is not available on any of the links on its route.

This model is a priority queue and a loss model. Variants and various aspects of trunk reservation models have been investigated by many authors, including Mitra, Gibbens, and Huang [8], [9], Mitra and Gibbens [7], Hunt and Laws [2], and Roberts [14], [15]. In [7], [8], [9] the authors considered symmetric loss networks with trunk

*Received by the editors July 23, 2007; accepted for publication (in revised form) October 27, 2008; published electronically February 25, 2009.

http://www.siam.org/journals/siap/69-5/69816.html

[†]Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (knessl@uic.edu). This author's work was partially supported by NSF grants DMS 02-02815 and DMS 05-03745 and by NSA grant H 98230-08-1-0102.

[‡]Consultant, Alcatel-Lucent Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974 (johnmorrison@alcatel-lucent.com).

reservation and dynamic routing. They analyze in detail the case of a single queue (or link) and use the results to obtain approximations for more complicated loss networks. Optimization and control policies for related models are considered in [2]. In [14] and [15] the author obtains analytical approximations to the blocking probabilities for the two customer types, based on a certain approximate recurrence. Having a thorough understanding of a single queue is important since then fixed point approximations can be used to study more general loss networks; see Kelly [3], [4].

The total load from the two customer types is defined as $\lambda+\nu/\kappa$. This represents the number of customers that would, on average, be in the system if the number of servers were *infinite*. If the total load exceeds the number of servers ($\lambda+\nu/\kappa > C$), then we are in the "overloaded" case, while if $\lambda+\nu/\kappa < C$, we are in the "underloaded" case. Important in applications is the case of "critical loading" where $\lambda+\nu/\kappa \approx C$. Here the system's capacity is roughly equal to the load.

We consider the asymptotic limit where $C \to \infty$ and the arrival rates λ and $\nu \to \infty$, at the same rate as C. Furthermore we assume that the number R of reserved trunks is $O(1)$ and that $C-(\lambda+\nu/\kappa) = O(\sqrt{C})$, so we are in the critical loading case.

We will pay particular attention to computing the blocking probabilities for both the primary and secondary traffic. In a loss model such as this, these give the probability that a phone call is lost on the link.

Previous asymptotic analyses of this model are due to Morrison [10] and the present authors [5], [12], [13]. However, in [10], [12], and [13] the analyses assume that the arrival rate of one traffic type is asymptotically greater than the other traffic type. This leads to some mathematical simplifications, namely, that one needs to solve parabolic rather than elliptic PDEs in the heavy traffic limit. Here we treat the more difficult case where the arrival rates of the two traffic types are comparably large. In [5] we considered the same model as in the present paper, with arrival rates of the same order, but for underloaded and overloaded links rather than a critically loaded one. This again led to much simpler mathematical problems.

We assume that both arrival rates and both service rates are of comparable magnitude, and we obtain a two-dimensional diffusion equation satisfied by the joint steady state probability distribution in this asymptotic limit. We derive this by using singular perturbation methods, and obtaining the appropriate boundary condition for this diffusion equation involves the analysis of the two different scales, i.e., ranges of (n_1, n_2). The analysis makes certain assumptions about not only the forms of various asymptotic expansions but also the asymptotic matching between the two scales. We then analyze the diffusion equation by a semi-analytic, seminumerical approach. This employs the classic technique of separation of variables, but satisfying the boundary condition is done numerically.

The paper is organized as follows. In section 2 we formulate the mathematical problem and introduce the asymptotic limit. In section 3 we obtain the diffusion equation and the boundary condition. We use analytic methods in section 4 to convert the problem into determining an infinite sequence of constants, and these we obtain numerically in section 5. In section 6 we give a brief summary.

2. Statement of the problem.

We let $N_1(t)$ (resp., $N_2(t)$) denote the number of servers occupied by primary (resp., secondary) customers. The steady state distribution will be denoted by

$$p(n_1, n_2) = \lim_{t\to\infty} \mathrm{Prob}[N_1(t) = n_1, N_2(t) = n_2]. \tag{2.1}$$

It satisfies the difference equation

$$\begin{aligned}[\lambda I\{n_1+n_2+1\leqslant C\}+\nu I\{n_1+n_2+1\leqslant C-R\}+n_1+\kappa n_2]p(n_1,n_2)\\ = \lambda I\{n_1\geqslant 1\}p(n_1-1,n_2)\\ +\nu I\{n_1+n_2\leqslant C-R\}I\{n_2\geqslant 1\}p(n_1,n_2-1)\\ +I\{n_1+n_2+1\leqslant C\}(n_1+1)p(n_1+1,n_2)\\ +\kappa I\{n_1+n_2+1\leqslant C\}I\{n_2+1\leqslant C-R\}(n_2+1)p(n_1,n_2+1)\end{aligned} \tag{2.2}$$

for $n_1 \geqslant 0$, $0 \leqslant n_2 \leqslant C-R$, $n_1+n_2 \leqslant C$. Here $I\{\mathcal{A}\}$ is the indicator function on the event $\mathcal{A}$.

We also have the normalization condition

$$\sum_{n_2=0}^{C-R}\sum_{n_1=0}^{C-n_2} p(n_1,n_2) = 1. \tag{2.3}$$

Of particular interest are the blocking probabilities, defined by

$$B_1 = \sum_{n_1=R}^{C} p(n_1, C-n_1), \quad B_2 = \sum_{\ell=0}^{R}\sum_{n_1=\ell}^{C-R+\ell} p(n_1, C-R+\ell-n_1). \tag{2.4}$$

The use of indicator functions allows us to write the problem as the single equation (2.2), and this form is useful for programming the numerical or symbolic solutions. But, the form (2.2) somewhat obscures the structure of the equation(s). We note that

$$\begin{aligned}(\lambda+n_1+\kappa n_2)p(n_1,n_2) = \lambda p(n_1-1,n_2)\\ +(n_1+1)p(n_1+1,n_2)\\ +\kappa(n_2+1)p(n_1,n_2+1)\end{aligned} \tag{2.5}$$

in the discrete oblique strip where $0 < n_2 < C-R$ and $C-R < n_1+n_2 < C$, and

$$\begin{aligned}(\lambda+\nu+n_1+\kappa n_2)p(n_1,n_2) = \lambda p(n_1-1,n_2)+\nu p(n_1,n_2-1)\\ +(n_1+1)p(n_1+1,n_2)\\ +\kappa(n_2+1)p(n_1,n_2+1)\end{aligned} \tag{2.6}$$

in the discrete triangle $0 < n_1+n_2 < C-R$. By analogy to PDEs we can view (2.5) as a parabolic problem (since the equation involves only first order differences in n_2), coupled to an elliptic problem (2.6), with an interface along $n_1+n_2 = C-R$. If $n_1+n_2 = C$, we have $(n_1+\kappa n_2)p(n_1,n_2) = \lambda p(n_1-1,n_2)$ which expresses the loss of primary customers. The loss of secondary customers is evident in (2.5) due to the absence of the term $\nu p(n_1,n_2-1)$.

We consider the asymptotic limit where $C,\lambda,\nu \to \infty$ with $R,\kappa = O(1)$. We introduce the new parameters σ and ρ, with

$$C-R = \frac{\lambda}{\kappa}\sigma, \quad \nu+\kappa\lambda = \rho\lambda. \tag{2.7}$$

Note that σ and ρ are to be $O(1)$. We consider only $p(n_1,n_2)$ for those ranges of n_1 and n_2, where most of the probability mass accumulates, in this asymptotic limit.

Now we introduce the heavy traffic assumption that $\rho \sim \sigma$, with δ defined by

$$\rho-\sigma = \frac{\delta}{\sqrt{\lambda}}, \tag{2.8}$$

where $\delta = O(1)$ and $-\infty < \delta < \infty$. Note that with (2.8) $C - R - (\lambda + \nu/\kappa) = O(\sqrt{\lambda})$ so that the total load $(\lambda + \nu/\kappa)$ is roughly balanced by the total number of servers. Most of the mass will occur where $n_1 \sim \lambda$ and $n_1 + n_2 \sim C$. More precisely, we introduce (x, y) with

$$n_1 = \lambda + x\sqrt{\lambda}, \quad n_2 = \left(\frac{\sigma}{\kappa} - 1\right)\lambda - (x + y)\sqrt{\lambda} \tag{2.9}$$

and note also that $n_1 + n_2 = C - R - y\sqrt{\lambda}$.

On the (x, y) scale we define

$$p(n_1, n_2) = \mathcal{P}(x, y; \lambda) \tag{2.10}$$

and for $y > 0$ we obtain from (2.2) or (2.6)

$$\begin{aligned}\Big\{\lambda(2 + 2\sigma - 2\kappa) + \sqrt{\lambda}\,[(1-\kappa)x - \kappa y + \delta]\Big\}&\mathcal{P}(x, y; \lambda)\\ = \lambda\mathcal{P}\left(x - \frac{1}{\sqrt{\lambda}}, y + \frac{1}{\sqrt{\lambda}}; \lambda\right)&\\ + \Big[(\sigma-\kappa)\lambda + \sqrt{\lambda}\delta\Big]\,\mathcal{P}\left(x, y + \frac{1}{\sqrt{\lambda}}; \lambda\right)&\\ + \Big(\lambda + \sqrt{\lambda}x + 1\Big)\,\mathcal{P}\left(x + \frac{1}{\sqrt{\lambda}}, y - \frac{1}{\sqrt{\lambda}}; \lambda\right)&\\ + \Big[(\sigma-\kappa)\lambda - \kappa\sqrt{\lambda}(x+y) + \kappa\Big]\,\mathcal{P}\left(x, y - \frac{1}{\sqrt{\lambda}}; \lambda\right).&\end{aligned} \tag{2.11}$$

In order to understand the boundary behavior of $\mathcal{P}$ near $y = 0$, we must also analyze (2.2) on another scale, namely, on the (x, ℓ) scale, where

$$n_1 + n_2 - (C - R) = \ell, \quad -\infty < \ell \leqslant R. \tag{2.12}$$

Note that the expansion for $y > 0$ may cease to be valid when y becomes small, and thus we include $\ell < 0$ in (2.12).

Then we set

$$p(n_1, n_2) = p_\ell(x) = \mathcal{P}(x, y; \lambda) \tag{2.13}$$

and from (2.2) we obtain

$$\begin{aligned}\Big[\lambda I\{\ell \leqslant R-1\} + [\lambda(\sigma-\kappa) + \delta\sqrt{\lambda}]I\{\ell \leqslant -1\}&\\ + \lambda - \lambda\kappa + \lambda\sigma + \sqrt{\lambda}(1-\kappa)x + \ell\kappa\Big]&p_\ell(x)\\ = \lambda p_{\ell-1}\left(x - \frac{1}{\sqrt{\lambda}}\right) + [\lambda(\sigma-\kappa) + \sqrt{\lambda}\delta]I\{\ell \leqslant 0\}p_{\ell-1}(x)&\\ + I\{\ell \leqslant R-1\}(\lambda + x\sqrt{\lambda} + 1)p_{\ell+1}\left(x + \frac{1}{\sqrt{\lambda}}\right)&\\ + I\{\ell \leqslant R-1\}\Big[\lambda(\sigma-\kappa) - \sqrt{\lambda}\kappa x + \kappa(\ell+1)\Big]\,p_{\ell+1}(x).&\end{aligned} \tag{2.14}$$

Here $\ell \leqslant R$ and we note that

$$\ell = -y\sqrt{\lambda}. \tag{2.15}$$

Note that some of the indicator functions in (2.2) were replaced by 1, since the local scaling (2.9) corresponds to $(n_1, n_2) \approx (\lambda, C-\lambda)$. However, the boundary $n_1+n_2 = C$ $(\ell = R)$ and interface $n_1+n_2 = C-R$ $(\ell = 0)$ do play key roles in the analysis. In the next section we analyze (2.11) and (2.14) asymptotically, ultimately obtaining a two-dimensional diffusion equation, with an appropriate boundary condition along $y = 0$.

3. Diffusion approximation. We rewrite (2.11) as

$$\begin{aligned}(3.1)\quad & \lambda\left[\mathcal{P}(x-\varepsilon, y+\varepsilon;\lambda) - \mathcal{P}(x,y;\lambda)\right] \\ &\qquad + \left[(\sigma-\kappa)\lambda + \delta\sqrt{\lambda}\right]\left[\mathcal{P}(x,y+\varepsilon;\lambda) - \mathcal{P}(x,y;\lambda)\right] \\ &+ \left(\lambda + \sqrt{\lambda}x\right)\left[\mathcal{P}(x+\varepsilon,y-\varepsilon;\lambda) - \mathcal{P}(x,y;\lambda)\right] + \mathcal{P}(x+\varepsilon,y-\varepsilon;\lambda) \\ &\quad + \left[(\sigma-\kappa)\lambda - \kappa(x+y)\sqrt{\lambda}\right]\left[\mathcal{P}(x,y-\varepsilon;\lambda) - \mathcal{P}(x,y;\lambda)\right] \\ &\qquad\qquad + \kappa\mathcal{P}(x,y-\varepsilon;\lambda) = 0,\end{aligned}$$

where $\varepsilon = 1/\sqrt{\lambda} \to 0^+$ and $y > 0$. We assume an expansion of the form

$$(3.2)\qquad \mathcal{P}(x,y;\lambda) = \frac{1}{\lambda}\left[\mathcal{P}(x,y) + \frac{1}{\sqrt{\lambda}}\mathcal{P}^{(1)}(x,y) + O\left(\lambda^{-1}\right)\right].$$

Here the first factor of $1/\lambda$ is suggested by the scaling (2.9), with which we expect that the normalizing sum (2.3) will be replaced asymptotically by a double integral over x and y.

Expanding (3.1) for $\varepsilon \to 0$ and using (3.2) we find that to leading order ($O(\varepsilon) = O(1/\sqrt{\lambda})$) the equation holds automatically, and at $O(\varepsilon^2) = O(1/\lambda)$ we obtain

$$(3.3)\quad \mathcal{P}_{xx} - 2\mathcal{P}_{xy} + (1+\sigma-\kappa)\mathcal{P}_{yy} + x\mathcal{P}_x + [\delta - x + \kappa(x+y)]\mathcal{P}_y + (\kappa+1)\mathcal{P} = 0.$$

This is a second order elliptic PDE that applies over the half-plane $y > 0, -\infty < x < \infty$. We note that $\sigma > \kappa$, since $\sigma \sim \rho$ and $\rho = \kappa + \nu/\lambda > \kappa$. If we change variables from (x, y) to (ξ, η) with

$$(3.4)\qquad x = \xi, \quad x + y = \eta, \quad \mathcal{P}(x,y) = P(\xi,\eta),$$

we obtain from (3.3)

$$(3.5)\qquad P_{\xi\xi} + (\sigma-\kappa)P_{\eta\eta} + \xi P_\xi + (\delta + \kappa\eta)P_\eta + (\kappa+1)P = 0,$$

which applies for $\eta > \xi$ and $-\infty < \xi < \infty$. Now, (3.5) is a separable PDE, but it applies over an oblique half-plane. At certain times we will use (3.5), while at other times (3.3) will prove more convenient.

We need a boundary condition along $y = 0$ for (3.3), or along $\xi = \eta$ for (3.5). To this end we must carefully consider the problem on the (x, ℓ) scale, where the interface condition(s) ($\ell = 0$ in (2.14)) play a role. We assume that on this scale we have the expansion

$$(3.6)\qquad p_\ell(x) = \frac{1}{\lambda}\left[p_\ell^{(0)}(x) + \frac{1}{\sqrt{\lambda}}p_\ell^{(1)}(x) + O(\lambda^{-1})\right].$$

For $R \geqslant 2$ and $1 \leqslant \ell \leqslant R-1$ we use (3.6) in (2.14). In this range $I\{\ell \leqslant -1\} = 0$ and $I\{\ell \leqslant R-1\} = 1$, and to leading order (2.14) gives

$$(3.7)\qquad (2+\sigma-\kappa)p_\ell^{(0)}(x) = p_{\ell-1}^{(0)}(x) + (1+\sigma-\kappa)p_{\ell+1}^{(0)}(x), \quad 1 \leqslant \ell \leqslant R-1.$$

When $\ell = R$, using (3.6) in (2.14) leads to

$$(1+\sigma-\kappa)p_R^{(0)}(x) = p_{R-1}^{(0)}(x). \tag{3.8}$$

Setting

$$a = \frac{1}{1+\sigma-\kappa} < 1, \tag{3.9}$$

the most general solution to (3.7) and (3.8) is

$$p_\ell^{(0)}(x) = a^\ell F_0(x), \quad 0 \leqslant \ell \leqslant R, \tag{3.10}$$

where F_0 is at this point undetermined.

At the next order of the expansion of (2.14) using (3.6) we obtain, for $1 \leqslant \ell \leqslant R-1$,

$$\begin{aligned}(2+\sigma-\kappa)p_\ell^{(1)}(x) &+ (1-\kappa)xp_\ell^{(0)}(x) \\ &= p_{\ell-1}^{(1)}(x) - \frac{d}{dx}p_{\ell-1}^{(0)}(x) + p_{\ell+1}^{(1)}(x) + \frac{d}{dx}p_{\ell+1}^{(0)}(x) + xp_{\ell+1}^{(0)}(x) \\ &\qquad + (\sigma-\kappa)p_{\ell+1}^{(1)}(x) - \kappa x p_{\ell+1}^{(0)}(x).\end{aligned} \tag{3.11}$$

By using (3.10) we can rearrange (3.11) to

$$\begin{aligned}p_{\ell-1}^{(1)}(x) - (2+\sigma-\kappa)p_\ell^{(1)}(x) &+ (1+\sigma-\kappa)p_{\ell+1}^{(1)}(x) \\ &= a^{\ell-1}(1-a)\left[a(1-\kappa)xF_0(x) + (1+a)F_0'(x)\right].\end{aligned} \tag{3.12}$$

The most general solution to the difference equation (3.12) is of the form

$$\begin{aligned}p_\ell^{(1)}(x) = a^\ell &\left\{p_0^{(1)}(x) - \ell\left[a(1-\kappa)xF_0(x) + (1+a)F_0'(x)\right]\right\} \\ &+ (a^\ell - 1)H_1(x), \quad 0 \leqslant \ell \leqslant R.\end{aligned} \tag{3.13}$$

Here we used $a(1+\sigma-\kappa) = 1$. When $\ell = R$, (2.14) with (3.6) yields at the second order

$$(1+\sigma-\kappa)p_R^{(1)}(x) + (1-\kappa)xp_R^{(0)}(x) = p_{R-1}^{(1)}(x) - \frac{d}{dx}p_{R-1}^{(0)}(x). \tag{3.14}$$

Using (3.10) and (3.13) in (3.14), we obtain

$$H_1(x) = -\frac{a^{R+1}}{1-a}F_0'(x). \tag{3.15}$$

We have thus obtained two terms in the expansion (3.6) for $\ell \geqslant 0$, up to the functions $F_0(x) = p_0^{(0)}(x)$ and $p_0^{(1)}(x)$. We have thus far assumed that $R \geqslant 2$, but we shall show that (3.10) holds if $R = 1$ also.

Now consider (2.14) for $R \geqslant 1$ and $\ell \leqslant 0$. In this case the expansion (3.6) leads to

$$\begin{aligned}\left[2+\sigma-\kappa+(\sigma-\kappa)I\{\ell \leqslant -1\}\right]&p_\ell^{(0)}(x) \\ &= (1+\sigma-\kappa)\left[p_{\ell-1}^{(0)}(x) + p_{\ell+1}^{(0)}(x)\right], \quad \ell \leqslant 0,\end{aligned} \tag{3.16}$$

so that

$$2p_\ell^{(0)}(x) = p_{\ell-1}^{(0)}(x) + p_{\ell+1}^{(0)}(x), \quad \ell \leqslant -1, \tag{3.17}$$

and

$$(2+\sigma-\kappa)p_0^{(0)}(x) = (1+\sigma-\kappa)\left[p_{-1}^{(0)}(x) + p_1^{(0)}(x)\right]. \tag{3.18}$$

The general solution to (3.17) is $F_0(x) + \ell G_0(x)$, where F_0 is as in (3.10). But then (3.18), with $p_1^{(0)} = aF_0$, shows that

$$p_{-1}^{(0)} = [(2+\sigma-\kappa)F_0 - F_0]\, a = F_0,$$

so that $G_0 = 0$ and hence

$$p_\ell^{(0)}(x) = F_0(x), \quad \ell \leqslant 0. \tag{3.19}$$

At the next order in using (3.6) in (2.14) we obtain, for $\ell \leqslant -1$,

$$\begin{aligned}(1+\sigma-\kappa)&\left[2p_\ell^{(1)}(x) - p_{\ell-1}^{(1)}(x) - p_{\ell+1}^{(1)}(x)\right] \\ &= \delta\left[p_{\ell-1}^{(0)}(x) - p_\ell^{(0)}(x)\right] + (1-\kappa)x\left[p_{\ell+1}^{(0)}(x) - p_\ell^{(0)}(x)\right] \\ &\qquad + \frac{d}{dx}\left[p_{\ell+1}^{(0)}(x) - p_{\ell-1}^{(0)}(x)\right],\end{aligned} \tag{3.20}$$

and $\ell = 0$ leads to

$$\begin{aligned}(2+\sigma-\kappa)p_0^{(1)}(x)&-(1+\sigma-\kappa)\left[p_{-1}^{(1)}(x)+p_1^{(1)}(x)\right] \\ &= (1-\kappa)x\left[p_1^{(0)}(x)-p_0^{(0)}(x)\right] + \delta p_{-1}^{(0)}(x) + \frac{d}{dx}\left[p_1^{(0)}(x)-p_{-1}^{(0)}(x)\right].\end{aligned} \tag{3.21}$$

In view of (3.19), the right-hand side of (3.20) vanishes so that

$$p_\ell^{(1)}(x) = F_1(x) + \ell G_1(x), \quad \ell \leqslant 0. \tag{3.22}$$

We use (3.10) and (3.13) with $\ell = 1$, (3.15), and $p_0^{(0)} = p_{-1}^{(0)} = F_0$ in (3.21). With (3.22), after some calculation this yields, for $R \geqslant 2$,

$$(1+\sigma-\kappa)G_1(x) = [\delta - (1-\kappa)x]F_0(x) - \left(2 - a^R\right)F_0'(x). \tag{3.23}$$

If $R = 1$ and $\ell = 1$, (2.14) yields to leading order

$$(1+\sigma-\kappa)p_1^{(0)}(x) = p_0^{(0)}(x) \tag{3.24}$$

and at the next order

$$(1+\sigma-\kappa)p_1^{(1)}(x) + (1-\kappa)xp_1^{(0)}(x) = p_0^{(1)}(x) - \frac{d}{dx}p_0^{(0)}(x). \tag{3.25}$$

But (3.24) shows that $p_1^{(0)}(x) = ap_0^{(0)}(x)$ so that (3.10) remains valid when $R = 1$, and (3.25) becomes

$$(1+\sigma-\kappa)p_1^{(1)}(x) = p_0^{(1)}(x) - F_0'(x) - a(1-\kappa)xF_0(x). \tag{3.26}$$

Using (3.13) and (3.15) we see that (3.26) is the same as (3.13) with $\ell = 1$, so (3.13) remains valid also when $R = 1$. It follows that (3.23) is also true if $R = 1$.

We now compare expansions (3.2) and (3.6). Setting $y = -\ell/\sqrt{\lambda}$ in (3.2) leads to

$$\frac{1}{\lambda}\left\{\mathcal{P}(x,0) + \frac{1}{\sqrt{\lambda}}\left[\mathcal{P}^{(1)}(x,0) - \ell\mathcal{P}_y(x,0)\right] + O(\lambda^{-1})\right\}. \tag{3.27}$$

If this is to agree with (3.6) for $\ell < 0$, we must have, in view of (3.19) and (3.22),

$$\mathcal{P}(x,0) = F_0(x), \tag{3.28}$$

$$\mathcal{P}^{(1)}(x,0) = F_1(x), \tag{3.29}$$

and

$$G_1(x) = -\mathcal{P}_y(x,0). \tag{3.30}$$

But then (3.23), (3.28), and (3.30) yield

$$(1+\sigma-\kappa)\mathcal{P}_y(x,0) = \left(2 - a^R\right)\mathcal{P}_x(x,0) + [(1-\kappa)x - \delta]\mathcal{P}(x,0). \tag{3.31}$$

This is the boundary condition that we sought for the leading term in (3.2). Calculating F_1 and $\mathcal{P}^{(1)}$ would require computing further terms in the expansions. In terms of (ξ, η), (3.31) becomes

$$(1+\sigma-\kappa)P_\eta(\xi,\xi) + \left(a^R - 2\right)[P_\xi(\xi,\xi) + P_\eta(\xi,\xi)] = [(1-\kappa)\xi - \delta]P(\xi,\xi). \tag{3.32}$$

To summarize, we have formulated the problem for the leading term in (3.2) as the PDE (3.3) and the boundary condition (BC) (3.31). Alternately, in terms of ξ and η we have the PDE (3.5) and BC (3.32). Using the Euler–MacLaurin formula we can show that the normalization condition in (2.3) becomes, to leading order,

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\mathcal{P}(x,y)\,dy\,dx = 1 \tag{3.33}$$

or

$$\int_{-\infty}^{\infty}\int_{\xi}^{\infty}P(\xi,\eta)\,d\eta\,d\xi = 1. \tag{3.34}$$

We note that the fact that $p(n_1, n_2)$ has a different form on the (x, ℓ) scale for $\ell > 0$ would affect the $O(1/\sqrt{\lambda})$ correction to the normalization.

Finally, we briefly discuss the cases $R = 0$ and $R \to \infty$. When $R = 0$ there are no reserved trunks, and then we have the exact product form solution

$$p(n_1,n_2) = C_0\frac{\lambda^{n_1}(\nu/\kappa)^{n_2}}{n_1!n_2!}, \tag{3.35}$$

where C_0 is a normalizing constant. By using in (3.35) the heavy traffic scaling in (2.8), and also using (2.7) and (2.9), we obtain after some calculation

$$p(n_1,n_2) \sim (\text{const.})e^{-x^2/2}\exp\left[-\frac{(\kappa(x+y)+\delta)^2}{2\kappa(\sigma-\kappa)}\right]. \tag{3.36}$$

While our derivation of (3.3) and (3.31) assumed that $R \geqslant 1$, a similar analysis shows that in fact the diffusion approximation holds also if $R = 0$. In this case we replace the coefficient $2 - a^R$ in (3.31) by 1. Then we can easily verify that (3.36) satisfies the PDE (3.3) and the BC (3.31). Thus our analysis is certainly consistent with the exact product form solution if $R = 0$.

Now consider $R \to \infty$ with σ and κ fixed, so that a is fixed and $C \to \infty$. If $R \to \infty$, we can replace $2 - a^R$ in (3.31) by 2, and (3.32) becomes

$$2P_\xi + \left(2 - \frac{1}{a}\right) P_\eta + [\xi(1-\kappa) - \delta]P = 0, \quad \xi = \eta. \tag{3.37}$$

We define

$$\widetilde{P}(\xi) = \int_\xi^\infty P(\xi, \eta)\, d\eta = \int_0^\infty \mathcal{P}(\xi, y)\, dy. \tag{3.38}$$

By integrating (3.3) from $y = 0$ to $y = \infty$, setting $x = \xi$, and using (3.31), we get

$$\widetilde{P}''(\xi) + \xi \widetilde{P}'(\xi) + \widetilde{P}(\xi) = 0 \tag{3.39}$$

so that

$$\widetilde{P}(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}. \tag{3.40}$$

Here we also used (3.34) for $R = \infty$. This shows that if $R = \infty$, the lowest order asymptotic approximation to the marginal density is exactly Gaussian. But, the structure of the two-dimensional density $P(\xi, \eta)$ seems much more complicated. The explicit solutions for $R = 0$ and $R = \infty$ can be used as useful checks on any numerical method for solving this boundary value problem.

4. Modal expansions. We consider (3.5) and (3.32). The PDE (3.5) admits solutions of the separable form

$$P(\xi, \eta) = f(\xi) g(\eta). \tag{4.1}$$

Using (4.1) in (3.5), dividing by fg, and rearranging terms yields

$$-\left[\frac{f''(\xi)}{f(\xi)} + \xi \frac{f'(\xi)}{f(\xi)} + 1\right] = (\sigma - \kappa)\frac{g''(\eta)}{g(\eta)} + (\delta + \kappa\eta)\frac{g'(\eta)}{g(\eta)} + \kappa. \tag{4.2}$$

Since the left-hand side of (4.2) depends only on ξ and the right-hand side only on η, both must be constant. We call this separation constant μ, and then

$$f''(\xi) + \xi f'(\xi) + (\mu + 1) f(\xi) = 0, \tag{4.3}$$

$$(\sigma - \kappa) g''(\eta) + (\delta + \kappa\eta) g'(\eta) + (\kappa - \mu) g(\eta) = 0. \tag{4.4}$$

Both of these ODEs are parabolic cylinder equations [6].

We expect P to have rapid decay as $|\xi|$ and/or $|\eta| \to \infty$ with $\eta \geqslant \xi$. By choosing $\mu = r = 0, 1, 2, \dots$ the solution to (4.3) has Gaussian decay as $\xi \to \pm\infty$, with

$$f(\xi) = e^{-\xi^2/4} D_r(\xi) = e^{-\xi^2/2} \mathrm{He}_r(\xi), \tag{4.5}$$

where He is the Hermite polynomial [6]. With $\mu = r$ we can require the solution of (4.4) to decay as $\eta \to +\infty$, and then

$$g(\eta) = \exp\left[-\frac{(\delta+\kappa\eta)^2}{4\kappa(\sigma-\kappa)}\right] D_{-r/\kappa}\left(\frac{\delta+\kappa\eta}{\sqrt{\kappa(\sigma-\kappa)}}\right). \tag{4.6}$$

We note that, since $\eta \geqslant \xi$, we have $\xi \to -\infty$ whenever $\eta \to -\infty$. By linear superposition and the completeness of the Hermite polynomials we argue that a general solution to (3.5) is

$$P(\xi,\eta) = e^{-\xi^2/4} \exp\left[-\frac{(\delta+\kappa\eta)^2}{4\kappa(\sigma-\kappa)}\right] \sum_{r=0}^{\infty} A(r) D_r(\xi) D_{-r/\kappa}\left(\frac{\delta+\kappa\eta}{\sqrt{\kappa(\sigma-\kappa)}}\right). \tag{4.7}$$

The constants $A(r)$ must be determined by the boundary condition (3.32). A crude but effective numerical method is to truncate the sum in (4.7) at some $r = N_{\max}$ and require that (3.32) hold at $\xi = \eta$ for some discrete set of points ξ_j. The totality of these points should suffice to determine $A(r)$ for $0 \leqslant r \leqslant N_{\max}$. We employ such a method in section 5.

Our choice of having $f(\xi)$ decay as $\xi \to \pm\infty$ was somewhat arbitrary, as we could also require that $g(\eta)$ have Gaussian decay as $\eta \to \pm\infty$. Then $g(\eta)$ would involve Hermite polynomials, which forces $\mu = -p\kappa$ for $p = 0, 1, 2, \ldots$, and thus

$$g(\eta) = \exp\left[-\frac{(\delta+\kappa\eta)^2}{4\kappa(\sigma-\kappa)}\right] D_p\left(\frac{\delta+\kappa\eta}{\sqrt{\kappa(\sigma-\kappa)}}\right). \tag{4.8}$$

The solution to (4.3) that decays as $\xi \to -\infty$ is

$$f(\xi) = e^{-\xi^2/4} D_{-p\kappa}(-\xi). \tag{4.9}$$

Again using linear superposition we argue that another form of the general solution to (3.5) is

$$P(\xi,\eta) = e^{-\xi^2/4} \exp\left[-\frac{(\delta+\kappa\eta)^2}{4\kappa(\sigma-\kappa)}\right] \sum_{p=0}^{\infty} C(p) D_{-p\kappa}(-\xi) D_p\left(\frac{\delta+\kappa\eta}{\sqrt{\kappa(\sigma-\kappa)}}\right). \tag{4.10}$$

Here $C(p)$ must be determined by the BC in (3.32).

We next discuss the normalization (3.34) and the blocking probabilities in (2.4). Let us define

$$\omega_0(y) = \int_{-\infty}^{\infty} \mathcal{P}(x,y)\,dx = \int_{-\infty}^{\infty} P(\xi, y+\xi)\,d\xi \tag{4.11}$$

so that the normalization (3.33) or (3.34) yields

$$\int_0^{\infty} \omega_0(y)\,dy = 1. \tag{4.12}$$

Using (2.4) with the scaling in (2.9) and (2.15), and also using (3.6), (3.10), and (3.28), we find that the blocking probabilities are, to leading order, given by

$$\begin{aligned} B_1 \sim \frac{1}{\sqrt{\lambda}} \int_{-\infty}^{\infty} p_R^{(0)}(x)\,dx &= \frac{a^R}{\sqrt{\lambda}} \int_{-\infty}^{\infty} \mathcal{P}(x,0)\,dx \\ &= \frac{1}{\sqrt{\lambda}} a^R \omega_0(0) \end{aligned} \tag{4.13}$$

and

$$B_2 \sim \frac{1}{\sqrt{\lambda}} \int_{-\infty}^{\infty} \sum_{\ell=0}^{R} p_\ell^{(0)}(x)\, dx = \frac{1-a^{R+1}}{\sqrt{\lambda}(1-a)} \int_{-\infty}^{\infty} \mathcal{P}(x,0)\, dx \tag{4.14}$$

$$= \frac{1}{\sqrt{\lambda}} \frac{1-a^{R+1}}{1-a} \omega_0(0).$$

Using properties of parabolic cylinder functions in [6], we show in the appendix that, for $r \geqslant 0$,

$$\int_0^\infty \!\! \int_{-\infty}^{\infty} e^{-x^2/4} D_r(x) \exp\left[-\frac{(\delta+\kappa(x+y))^2}{4\kappa(\sigma-\kappa)}\right] D_{-r/\kappa}\left(\frac{\delta+\kappa(x+y)}{\sqrt{\kappa(\sigma-\kappa)}}\right) dx\, dy \tag{4.15}$$

$$= \sqrt{2\pi}(-1)^r \left(\sqrt{\frac{\kappa}{\sigma}}\right)^{r-1} \left(\sqrt{1-\frac{\kappa}{\sigma}}\right)^{1-r/\kappa} \exp\left(-\frac{\delta^2}{4\sigma\kappa}\right) D_{r-1-r/k}\left(\frac{\delta}{\sqrt{\sigma\kappa}}\right).$$

Then if we define $\alpha(r)$ from

$$\alpha(r) = \sqrt{2\pi}(-1)^r \left(\sqrt{\frac{\kappa}{\sigma}}\right)^{r} \left(\sqrt{1-\frac{\kappa}{\sigma}}\right)^{1-r/\kappa} \exp\left(-\frac{\delta^2}{4\sigma\kappa}\right) A(r) \tag{4.16}$$

and use the expansion (4.7) for $P(\xi,\eta) = \mathcal{P}(x, x+y)$, the normalization (4.12) becomes

$$\sqrt{\frac{\kappa}{\sigma}} = \sum_{r=0}^{\infty} \alpha(r) D_{r-1-r/\kappa}\left(\frac{\delta}{\sqrt{\sigma\kappa}}\right). \tag{4.17}$$

Similarly we show in the appendix that

$$\omega_0(0) = \sum_{r=0}^{\infty} \alpha(r) D_{r-r/\kappa}\left(\frac{\delta}{\sqrt{\sigma\kappa}}\right). \tag{4.18}$$

If instead of using (4.7) we use (4.10), then the analogous results are

$$\sqrt{\frac{\kappa}{\sigma}} = \sum_{p=0}^{\infty} \gamma(p) D_{p-1-p\kappa}\left(\frac{\delta}{\sqrt{\sigma\kappa}}\right), \tag{4.19}$$

where

$$\gamma(p) = \sqrt{2\pi} \left(\sqrt{\frac{\sigma}{\kappa}}\right)^{p\kappa} \left(\sqrt{1-\frac{\kappa}{\sigma}}\right)^{p+1} \exp\left(-\frac{\delta^2}{4\sigma\kappa}\right) C(p) \tag{4.20}$$

and

$$\omega_0(0) = \sum_{p=0}^{\infty} \gamma(p) D_{p-p\kappa}\left(\frac{\delta}{\sqrt{\sigma\kappa}}\right). \tag{4.21}$$

We comment that if $R = 0$, retaining only the $r = 0$ term in (4.7) or the $p = 0$ term in (4.10) regains the exact solution in (3.36).

5. Numerical studies. We consider now the numerical solution of (3.5), with (3.32) and (3.34). We use the C-mode expansion in (4.10), which proved numerically superior to the A-mode expansion in (4.7).

The basic method is as follows. We impose on (4.10) (or (4.7)) the boundary condition (3.32). First, we truncate the sum in (4.10) at $p = N_{\max}$ (or (4.7) at $r = N_{\max}$) so we approximate $P(\xi, \eta)$ using the first $N_{\max} + 1$ modes. Then we require (3.32) to hold along $\eta = \xi = u$ for certain discrete points ξ_j. For example, we choose some interval $[u_0 - AA, u_0 + AA]$ centered at u_0 and require (3.32) to hold at the points

$$u = u_0 + AA\frac{N}{N_0}; \quad N = -N_0, -N_0 + 1, \ldots, -2, -1, 1, 2, \ldots, N_0 - 1, N_0. \tag{5.1}$$

Here we omit $u = u_0$ and instead impose the normalization (3.34) in the form (4.20) (or (4.17)), with the sum again truncated at $p = N_{\max}$ (or $r = N_{\max}$). Using (5.1) yields $2N_0$ equations, and thus the $N_{\max} + 1$ coefficients in (4.10) satisfy $2N_0 + 1$ equations. If $N_{\max} = 2N_0$, we obtain an inhomogeneous linear system that should uniquely determine (an approximation to) the $C(p)$ for $0 \leqslant p \leqslant N_{\max}$. But, we have found that this approach leads to numerical instabilities. Instead we use an approach similar to that in [11] based on the method of least squares. Here we take a relatively small $N_{\max}$ and a much larger N_0, so that the points in (5.1) densely fill the interval $u \in [u_0 - AA, u_0 + AA]$, and obtain the least squares approximate solution to the linear system. Then we increase N_0 to get convergence of the first few $C(p)$ as $N_0 \to \infty$ for this fixed value of $N_{\max}$. We increase $N_{\max}$ and repeat the procedure. Thus if we write $C(p) = C(p; N_0, N_{\max})$, we first get convergence as $N_0 \to \infty$, and then as $N_{\max} \to \infty$, for the first few coefficients in (4.10).

Our studies show that $P(u, u)$ is a unimodal function with very thin tails, which corresponds to the boundary values of the two-dimensional density $P(\xi, \eta)$. We choose u_0 to be close to where $P(u, u)$ is peaked, and we have found that taking $AA = 5$ is more than sufficient to capture the left and right tails of $P(u, u)$. We have found that a good way of checking the numerical convergence of our method is to plot $P(u, u)$ for a given $N_{\max}$, and then for $N_{\max} + 5$ (using 5 additional modes), and see if the graphs coincide.

To illustrate the numerical method we first take $a = .5$, $\kappa = 2$, $\delta = .01$, and $R = 3$ (thus $\sigma = 3$). In Table 1 we used $u_0 = 1$ and $AA = 5$, and we give the first three coefficients $(C(0), C(1), C(2))$ in (4.10) for various values of $N_{\max}$. We start with $N_{\max} = 4$ (5 modes) and increase this in increments of 5. We also give the values of B_1^0 and B_2^0, where

$$B_1^0 = a^R \omega_0(0), \quad B_2^0 = \frac{1 - a^{R+1}}{1 - a}\omega_0(0), \tag{5.2}$$

and $\omega_0(0)$ is computed using (4.21) and (4.20), with the sum truncated at $p = N_{\max}$. It follows that $B_j^0/\sqrt{\lambda}$ should be an approximation to the blocking probabilities B_j for $j = 1, 2$. For each $N_{\max}$ we increased N_0 until we got convergence of $C(p)$ and B_j^0. We started with $N_{\max} = 4$ and $N_0 = 35$, so we used a least squares approximation to a system of 71 equations in 5 unknowns.

In Table 1 we consider 5, 10, 15, 20, and 25 modes. We see that the convergence of $C(0)$ with increasing $N_{\max}$ is quite rapid, but the coefficient $C(2)$ of the third mode is converging much more slowly. However, the B_j^0 converge very rapidly. In Figure 1 we plot the curve $P(u, u)$ for $-4 \leqslant u \leqslant 6$. Here we used $N_{\max} + 1$ modes in (4.10)

Table 1

$\delta = .01$, $a = .5$, $\kappa = 2$, $R = 3$.

$N_{\max}$	$C(0)$	$C(1)$	$C(2)$	B_1^0	B_2^0
4	.261	.272	$-.0328$	.102	1.53
9	.248	.301	$-.115$	.103	1.55
14	.246	.311	$-.164$	.103	1.55
19	.246	.315	$-.189$	.103	1.55
24	.245	.316	$-.201$	.103	1.55

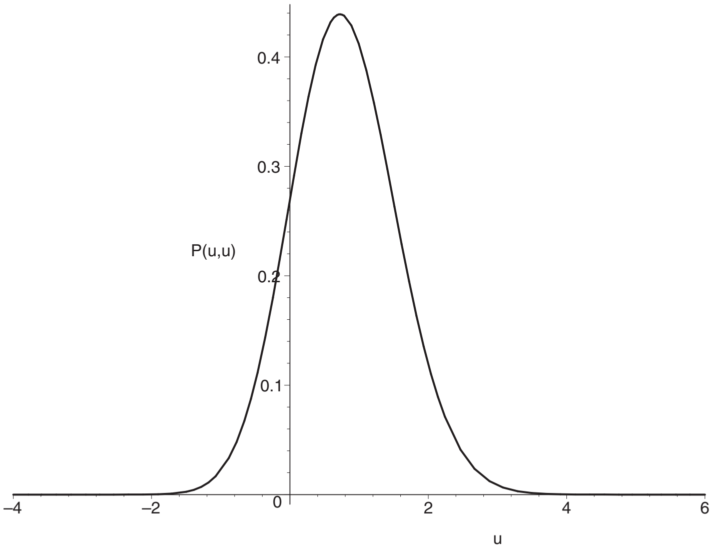


Fig. 1. *A sketch of the curve $P(u,u)$ for $-4 \leq u \leq 6$ when $\delta = .01$.*

and set $\xi = \eta = u$. Like the B_j^0 the graph converges quite rapidly with $N_{\max}$, and we get indistinguishable curves using 15 or 25 modes. The graph shows that $P(u,u)$ has very thin tails. Our results also show that the maximum value of $P(u,u)$ is about .438 and this occurs at $u \cong .723$. In Figures 2 and 3 we plot the "surface" $P(u,v)$ obtained from (4.10) with the sum truncated at $p = N_{\max}$. This also is quite robust and does not change much with increasing $N_{\max}$. In Figure 2 we view the surface from the direction $u = v > 0$ in the (u,v) plane, while Figure 3 uses $\Theta = -15°$, where $u = \sqrt{u^2+v^2}\cos(\Theta)$, $v = \sqrt{u^2+v^2}\sin(\Theta)$.

In Table 2, and Figures 4, 5, and 6 we use $\delta = -3.5$, retaining the other parameter values. Note that as $\delta \to -\infty$ we are moving out of the critically loaded case into the underloaded case (we have, by (2.7) and (2.8), $C - R = \lambda + \nu/\kappa - \sqrt{\lambda}\delta/\kappa$). We use $u_0 = 2$ and plot $P(u,u)$ for $u \in (-3,7)$. Figure 4 shows that decreasing δ from .01 to -3.5 tends to make the maximum of $P(u,u)$ much smaller ($< .09$) and shifts the peak to the right. The surface $P(u,v)$ is given in Figures 5 and 6, from the same two perspectives as in Figures 2 and 3. We see that now $P(u,v)$ has an interior maximum in the range $v > u$. From Table 2 we see that the blocking coefficients B_j^0 are now smaller, which is related to the fact that there is less mass near the boundary $u = v$. Table 2 also shows that the coefficient $C(0)$ associated with the first C-mode has a much larger numerical value than $C(1)$ or $C(2)$, suggesting that retaining only the $p = 0$ term in (4.10) may yield a reasonable approximation. For $\delta \to -\infty$ we can give an asymptotic argument, by using asymptotic matching to the underloaded case, that the solution to (3.5) should be a product of two Gaussians, as in (3.36).

Table 2

$\delta = -3.5$, $a = .5$, $\kappa = 2$, $R = 3$.

$N_{\max}$	$C(0)$	$C(1)$	$C(2)$	B_1^0	B_2^0
4	.236	.00990	−.00667	.0165	.248
9	.229	.0171	−.0152	.0178	.268
14	.228	.0200	−.0216	.0179	.269
19	.227	.0214	−.0253	.0181	.272
24	.227	.0218	−.0273	.0181	.271

Table 3

$\delta = 3.5$, $a = .5$, $\kappa = 2$, $R = 3$.

$N_{\max}$	$C(0)$	$C(1)$	$C(2)$	B_1^0	B_2^0
4	.623	1.54	9.44	.244	3.67
9	.312	2.17	4.96	.258	3.87
14	.313	2.14	5.24	.258	3.87
19	.312	2.14	5.17	.258	3.87
24	.312	2.14	5.15	.258	3.87

Table 4

$\delta = .01$, $a = .5$, $\kappa = 2$, $R = 3$.

C	λ	ν	$\sqrt{\lambda}B_1$	B_1^0	$\sqrt{\lambda}B_2$	B_2^0
10	4.666	4.688	.0470	.103	.884	1.55
20	11.33	11.36	.0719	.103	1.07	1.55
30	18.00	18.04	.0812	.103	1.15	1.55
40	24.66	24.71	.0861	.103	1.20	1.55
50	31.33	31.38	.0891	.103	1.24	1.55
60	38.00	38.06	.0912	.103	1.27	1.55
70	44.66	44.73	.0926	.103	1.29	1.55

Next we consider a fairly large positive value of δ. Table 3 and Figures 7, 8, and 9, have $\delta = +3.5$ (again with $a = .5$, $\kappa = 2$, $R = 3$). Compared to the other two cases, now the coefficients $C(p)$ increase in value with p. The convergence as $N_{\max}$ increases is similar to the other cases, and the blocking coefficients B_j^0 converge quickly. In the figures we use $u_0 = 0$ and $-5 \leqslant u, v \leqslant 5$. Now $P(u,u)$ has the maximum value of about .994 at $u \cong .227$. Figures 8 and 9 show that the two-dimensional density is now quite concentrated near the boundary $u = v$. Note that $\delta \to +\infty$ corresponds to going from the critically loaded case to the overloaded case. For the latter we showed in [5] that $p(n_1, n_2)$ is concentrated, where $\ell = n_1 + n_2 - (C - R) = O(1)$.

Up to now we have discussed the numerical solution of the diffusion equation, but not how accurate an approximation it is for the original discrete model. In Table 4 we consider the exact numerical values of the blocking probabilities, as computed by solving (2.2) and (2.3), and compare these to those obtained from the diffusion approximation. We compare $\sqrt{\lambda}B_j$ (cf. (2.4)) to B_j^0 (cf. (5.2)). To compute B_j^0 we need to input the parameters a, κ, R, and δ. From these we can get λ and ν from

$$\lambda = \frac{\kappa}{\sigma}(C - R) = \frac{\kappa}{a^{-1} - 1 + \kappa}(C - R), \tag{5.3}$$

$$\rho = \sigma + \frac{\delta}{\sqrt{\lambda}}, \tag{5.4}$$

and

$$\nu = (\rho - \kappa)\lambda = \frac{a^{-1} - 1 + \delta/\sqrt{\lambda}}{a^{-1} - 1 + \kappa}\kappa(C - R). \tag{5.5}$$

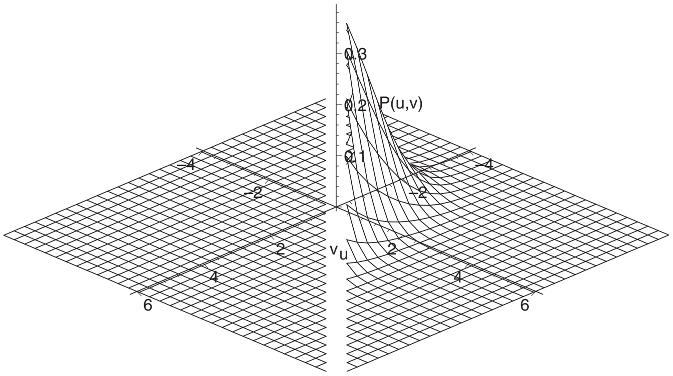


Fig. 2. *A sketch of the surface $P(u,v)$ from the direction $u = v$ when $\delta = .01$.*

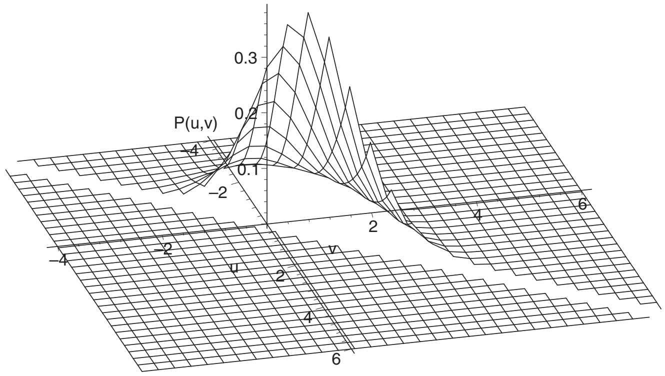


Fig. 3. *A sketch of the surface $P(u,v)$ from a different perspective when $\delta = .01$.*

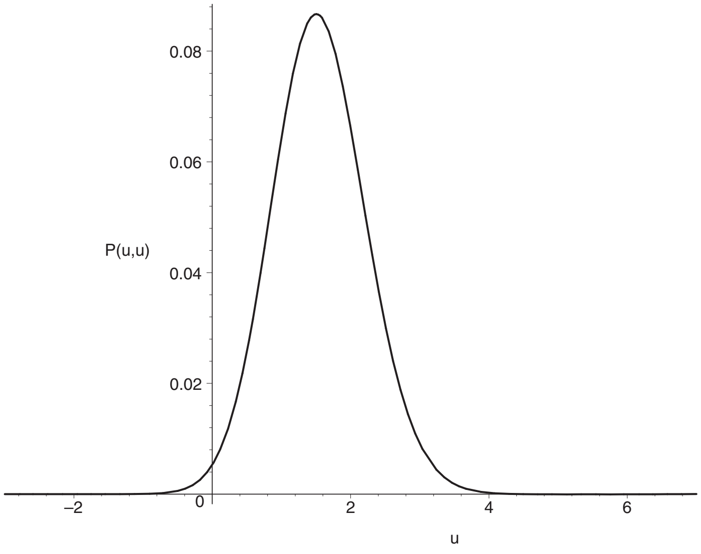


Fig. 4. *A sketch of the curve $P(u,u)$ for $-3 \leq u \leq 7$ when $\delta = -3.5$.*

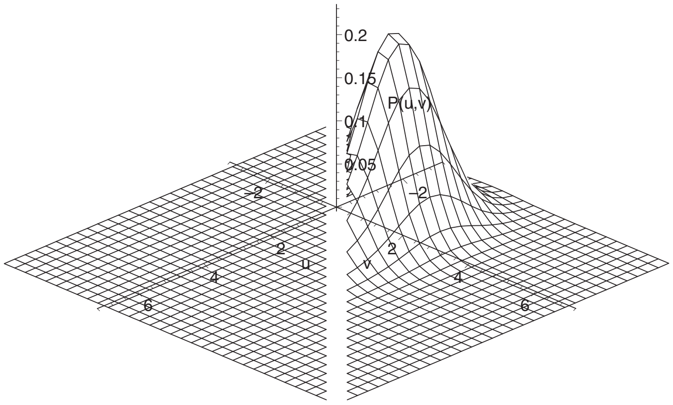


FIG. 5. *A sketch of the surface $P(u,v)$ from the direction $u = v$ when $\delta = -3.5$.*

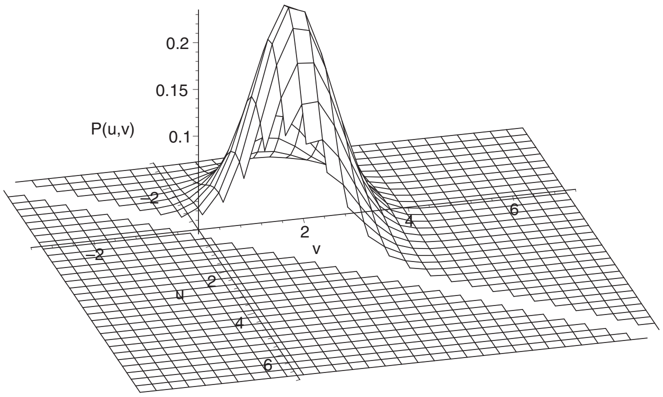


FIG. 6. *A sketch of the surface $P(u,v)$ from a different perspective when $\delta = -3.5$.*

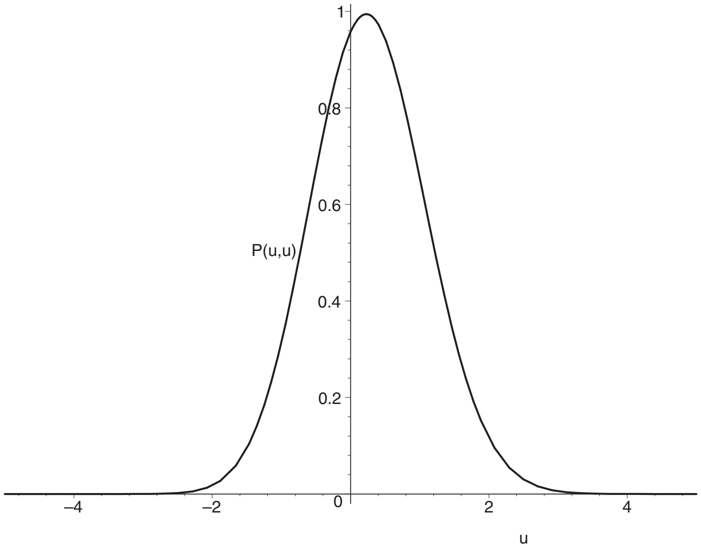


FIG. 7. *A sketch of the curve $P(u,u)$ for $-5 \leq u \leq 5$ when $\delta = 3.5$.*

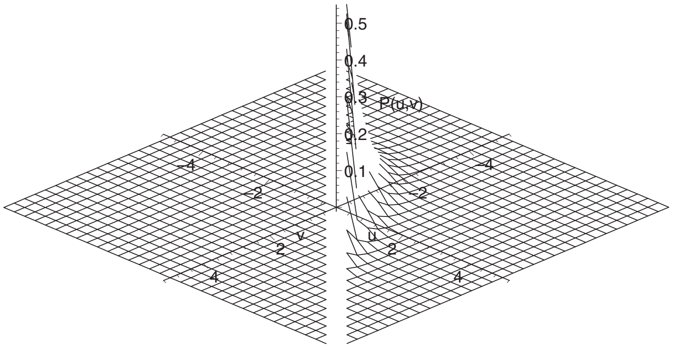


FIG. 8. *A sketch of the surface $P(u,v)$ from the direction $u=v$ when $\delta=3.5$.*

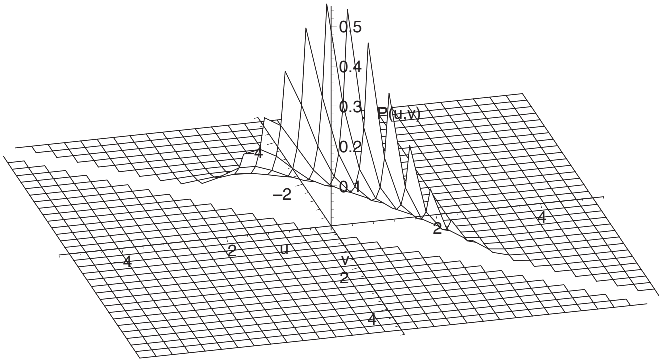


FIG. 9. *A sketch of the surface $P(u,v)$ from a different perspective when $\delta=3.5$.*

Then we input κ, R, C and λ,ν to solve the discrete linear system (2.2) numerically. Table 4 compares, for various C, the numerical values of $\sqrt{\lambda}B_j$ to B_j^0 (which are independent of C and λ). We see that $\sqrt{\lambda}B_j$ do appear to be converging as $C\to\infty$, but fairly slowly. The results are certainly consistent with the $O(1/\sqrt{\lambda})$ correction terms in, e.g., (3.2). We could in theory obtain the problem satisfied by the correction term to the diffusion approximation and analyze it numerically.

To summarize, the numerical method based on the C-modes does efficiently yield the blocking coefficients B_j^0, and determines the density $P(u,v)$, including $P(u,u)$, for ranges where there is significant mass. The method does have some problems with the tails, but here an asymptotic approach may be more appropriate.

6. Conclusion.

To summarize, we have analyzed a trunk reservation model in the limit of rapid arrivals for both high and low priority customer classes, a large number of servers or circuits, and a critical load condition in which the arrivals are roughly balanced by the number of circuits. Using a semi-analytic and seminumerical approach we derived and analyzed a two-dimensional PDE that describes the joint distribution of the numbers of circuits in use by both the high and low priority customers. Particular attention was paid to the blocking probabilities, which are both $O(1/\sqrt{\lambda})$ in this asymptotic limit.

This model may represent a single node in a circuit-switched network. However, the analysis of such a network depends on its topology, the capacities of its links and the offered traffic rates, and holding times between different pairs of origin and destination nodes. Consequently, it is not possible to say what our asymptotic results imply about the behavior of the network without a numerical analysis using fixed point approximations. This is beyond the scope of our paper.

Appendix. We will derive (4.17)–(4.19), using properties of the parabolic cylinder functions [6]. First, we use the property

$$e^{-z^2/4}D_\mu(z) = -\frac{d}{dz}\left[e^{-z^2/4}D_{\mu-1}(z)\right], \tag{A.1}$$

and note that $D_\nu(z) \to 0$ as $z \to \infty$, to obtain

$$\begin{aligned}\int_0^\infty \exp\left[-\frac{(\delta+\kappa(x+y))^2}{4\kappa(\sigma-\kappa)}\right] D_\mu\left(\frac{\delta+\kappa(x+y)}{\sqrt{\kappa(\sigma-\kappa)}}\right) dy \\ = \sqrt{\frac{\sigma}{\kappa}-1}\exp\left[-\frac{(\delta+\kappa x)^2}{4\kappa(\sigma-\kappa)}\right] D_{\mu-1}\left(\frac{\delta+\kappa x}{\sqrt{\kappa(\sigma-\kappa)}}\right).\end{aligned} \tag{A.2}$$

To evaluate $\omega_0(0)$ and the normalization condition (4.12) from (4.7) and (4.11), it will suffice, in view of (A.2), to consider the definite integral

$$I_m = \int_{-\infty}^\infty e^{-x^2/4}D_m(x)\exp\left[-\frac{1}{4}(a+bx)^2\right] D_{-\rho}(a+bx)\,dx \tag{A.3}$$

for $b > 0$ and m a nonnegative integer.

The term corresponding to $p = 0$ in the expansion (4.10) is equivalent to the term corresponding to $r = 0$ in (4.7). Hence we may restrict our attention to $p > 0$ in (4.10). If we let $a + bx = -\xi$ in (A.3), we obtain

$$\int_{-\infty}^\infty e^{-\xi^2/4}D_{-\rho}(-\xi)\exp\left[-\frac{(a+\xi)^2}{4b^2}\right] D_m\left(\frac{a+\xi}{b}\right) d\xi = (-1)^m bI_m. \tag{A.4}$$

Hence it will suffice to consider I_m for the expansion (4.10) also. But,

$$e^{-z^2/4}\sum_{m=0}^\infty D_m(z)\frac{t^m}{m!} = \exp\left[-\frac{1}{2}(z-t)^2\right]. \tag{A.5}$$

Hence, from (A.4), after some algebra, it follows that

$$\begin{aligned}\sum_{m=0}^\infty (-1)^m I_m\frac{t^m}{m!} = \frac{1}{b}\exp\left[-\frac{(a-bt)^2}{2(1+b^2)}\right] \\ \cdot\int_{-\infty}^\infty e^{\xi^2/4}D_{-\rho}(\xi)\exp\left\{-\frac{1+b^2}{2b^2}\left[\xi-\frac{a-bt}{1+b^2}\right]^2\right\} d\xi,\end{aligned} \tag{A.6}$$

where we have replaced ξ by $-\xi$.

Now by page 886 in [1],

$$\begin{aligned}\int_{-\infty}^\infty \exp\left[-\frac{(x-y)^2}{2\mu}\right] e^{x^2/4}D_\nu(x)\,dx \\ = \sqrt{2\pi\mu}(\sqrt{1-\mu})^\nu \exp\left[\frac{y^2}{4(1-\mu)}\right] D_\nu\left(\frac{y}{\sqrt{1-\mu}}\right), \quad 0 < \mu < 1.\end{aligned} \tag{A.7}$$

(There is a factor $\sqrt{2\pi\lambda}$ missing from the corresponding result in [6].) From (A.6) and (A.7), with $\mu = b^2/(1+b^2)$ and $y = (a-bt)/(1+b^2)$, we obtain

$$\text{(A.8)} \qquad \sum_{m=0}^{\infty}(-1)^m I_m \frac{t^m}{m!} = \frac{\sqrt{2\pi}}{(\sqrt{1+b^2})^{1-\rho}} \exp\left[-\frac{(a-bt)^2}{4(1+b^2)}\right] D_{-\rho}\left(\frac{a-bt}{\sqrt{1+b^2}}\right).$$

But,

$$\text{(A.9)} \qquad D_\nu(x+y) = \exp\left(\frac{xy}{2}+\frac{y^2}{4}\right)\sum_{m=0}^{\infty}\frac{(-y)^m}{m!}D_{m+\nu}(x).$$

It follows from (A.8) and (A.9) that

$$\text{(A.10)} \qquad I_m = \frac{\sqrt{2\pi}(-b)^m}{(\sqrt{1+b^2})^{m+1-\rho}} \exp\left[-\frac{a^2}{4(1+b^2)}\right] D_{m-\rho}\left(\frac{a}{\sqrt{1+b^2}}\right).$$

In (A.3), for the expansion (4.7), we take $a = \delta/\sqrt{\kappa(\sigma-\kappa)}$, $b = \sqrt{\kappa/(\sigma-\kappa)}$, and $m = r \geqslant 0$, and we take $\rho = 1 + r/\kappa$ for the evaluation of the normalization condition (4.12) and $\rho = r/\kappa$ for the evaluation of $\omega_0(0)$. The former choice of ρ, with the help of (A.2) and (A.10), leads to (4.15), and hence to (4.17), where $\alpha(r)$ is defined by (4.16). The latter choice of ρ leads to (4.18). For the expansion (4.10) we use (A.4), and in (A.3) we take $a = \delta/\kappa$, $b = \sqrt{(\sigma-\kappa)/\kappa}$, and $\rho = p\kappa$, and we take $m = p-1$, $p > 0$ for the evaluation of the normalization condition (4.12) and $m = p \geqslant 0$ for the evaluation of $\omega_0(0)$. The former choice of m, with the help of (A.2) and (A.10), leads to (4.19), where $\gamma(p)$ is defined by (4.20). The term corresponding to $p = 0$ in (4.10) is equivalent to that corresponding to $r = 0$ in (4.7), as pointed out previously. The latter choice of m leads to (4.21).

REFERENCES

[1] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 4th ed., Academic Press, New York, 1965.

[2] P. J. Hunt and C. N. Laws, *Optimization via trunk reservation in single resource loss systems under heavy traffic*, Ann. Appl. Probab., 7 (1997), pp. 1058–1079.

[3] F. P. Kelly, *Blocking probabilities in large circuit switched networks*, Adv. in Appl. Probab., 18 (1986), pp. 473–505.

[4] F. P. Kelly, *Loss networks*, Ann. Appl. Probab., 1 (1991), pp. 319–378.

[5] C. Knessl and J. A. Morrison, *Blocking probabilities for an underloaded or overloaded link with trunk reservation*, SIAM J. Appl. Math., 66 (2005), pp. 82–97.

[6] W. Magnus, F. Oberhettinger, and R. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, 3rd ed., Springer-Verlag, New York, 1966.

[7] D. Mitra and R. J. Gibbens, *State-dependent routing on symmetric loss networks with trunk reservations.* II. *Asymptotics, optimal design*, Ann. Oper. Res., 35 (1992), pp. 3–30.

[8] D. Mitra, R. J. Gibbens, and B. D. Huang, *Analysis and optimal design of aggregated-least-busy-alternative routing on symmetric loss networks with trunk reservations*, in Teletraffic and Datatraffic in a Period of Change, Proceedings of ITC-13, A. Jensen and V. B. Iversen, eds., North–Holland, Amsterdam, 1991, pp. 477–482.

[9] D. Mitra, R. J. Gibbens, and B. D. Huang, *State-dependent routing on symmetric loss networks with trunk reservations*, I, IEEE Trans. Comm., 41 (1993), pp. 400–411.

[10] J. A. Morrison, *Blocking probabilities for a single link with trunk reservation*, J. Math. Anal. Appl., 203 (1996), pp. 401–434.

[11] J. A. Morrison and M.-J. Cross, *Scattering of a plane electromagnetic wave by axisymmetric raindrops*, Bell System Tech. J., 53 (1974), pp. 955–1019.

[12] J. A. Morrison and C. Knessl, *Asymptotic analysis of a loss model with trunk reservation.* I. *Trunks reserved for fast traffic*, J. Appl. Math. Stoch. Anal., 2008, article 415692.

[13] J. A. Morrison and C. Knessl, *Asymptotic analysis of a loss model with trunk reservation. II. Trunks reserved for slow traffic*, Stud. Appl. Math., to appear.

[14] J. W. Roberts, *A service system with heterogeneous server requirements*, in Performance of Data Communications Systems and Their Applications, G. Pujolle, ed., North–Holland, Amsterdam, 1981, pp. 423–431.

[15] J. W. Roberts, *Teletraffic models for the Telecom I integrated services network*, in Proceedings of the 10th International Teletraffic Congress (Montreal), 1983, paper 1.1-2.

SIAM J. APPL. MATH.
Vol. 69, No. 5, pp. 1477–1500

NUMERICAL MODELING OF ELECTROWETTING BY A SHAPE INVERSE APPROACH*

JÉRÔME MONNIER†, PATRICK WITOMSKI†, PATRICK CHOW-WING-BOM†, AND CLAIRE SCHEID†

Abstract. We model an electrified droplet spreading on a solid surface. The model aims to seek a drop shape that minimizes its total energy (capillary, electrostatic, and gravitational). We derive the equations and the shape gradient; then we detail the shape optimization algorithm and present some numerical results. Up to a critical applied voltage value, the computed angles fit the predictions of Lippman's equation (plane capacitor approximation). Then, when increasing the voltage, we observe an overestimate of the Lippman prediction. Numerical computations of the curvature show that it remains constant everywhere except in the vicinity of the contact point, where it increases sharply.

Key words. shape optimal, design, electrowetting, contact point, energy minimization, curvature

AMS subject classifications. 49Q10, 74M15

DOI. 10.1137/050624340

1. Introduction.

Electrowetting can be defined as a tool for spreading liquid droplets (e.g., water) on hydrophobic solid surfaces (e.g., polymer film). This is quite a recent technique (see [1]) which holds very attractive properties for manipulation of tiny liquid volumes, as is done, for example, in biotechnologies. The principle of electrowetting is to apply an electric field between the conductor liquid droplet and the solid surface in order to change the droplet spreading on the surface. Given the liquid volume, the main feature for describing the droplet is the wetting angle.

Several articles discuss the experimental aspects of electrowetting and present some analytical analysis; see, e.g., [1], [21], [2], and the references therein. One property of electrowetting still poorly understood by physicists is the contact angle saturation. Several mechanisms for explaining it were proposed in [21], [22], [17], [20]. When increasing the applied electric voltage, the liquid droplet spreads onto the solid and the wetting angle decreases. Nevertheless, this is true only if the value of the applied voltage is less than a certain critical value. Up to this critical value, the contact angle can be derived from the Lippman equation using the plane capacitor approximation. For higher values, one observes a saturation of the wetting angle and for even higher values, instabilities of the contact line liquid-solid-gas can appear. A few hypotheses have been made to explain the saturation phenomenon. Let us cite, for example, the air ionization (see [21]) or electrostatic effects near the wetting line (see [4]). This limiting phenomenon is still under investigation and the full modeling of electrowetting remains an open problem. In other respects, the authors of [5] show that the contact angle does not depend on the potential. It remains equal to the static Young angle (obtained when the potential is null). Also, they observe that the curvature near the contact line increases while the potential increases.

*Received by the editors February 14, 2008; accepted for publication (in revised form) November 6, 2008; published electronically February 25, 2009.

http://www.siam.org/journals/siap/69-5/62434.html

†University of Grenoble and INRIA, Laboratory LJK (Moise project-team), 38041 Grenoble cedex 9, France (monnier@imag.fr, patrick.witomski@imag.fr, chow-wing-bom@imag.fr, claire.scheid@imag.fr).

In this study, we present a mathematical approach for modeling and numerically computing the drop shape, given an applied voltage. The model is based on the shape optimal design methods; see, e.g., [6], [11]. We seek the drop shape (a free surface) such that it minimizes its total energy. The total energy is the sum of the capillary energy, the gravitational energy, and the electrostatic energy. Our numerical modeling is general in the sense that we do not make any assumption on the drop shape. The equations are fully solved and the shape is defined in a general family of surfaces. We assume that the drop shape is steady state and remains two-dimensional (2D) axisymmetric but the method remains valid for three-dimensional (3D) shapes as well. Of course, in the 3D case, the implementation is much more complex and time-consuming than in the present 2D axisymmetric case. This 2D axisymmetric assumption is valid for applied voltages up to the value leading to the instabilities mentioned above.

We obtain numerical results which are consistent with the plane capacitor approximation (Lippman's equation) only for low voltages. For higher voltages, we observe an overestimate of the Lippman predictions. Nevertheless, with the present model, we do not retrieve the wetting angle saturation but instead a deviation from Lippman's predictions of the shape of the drop. In other respects, we focus on the curvature values of the droplet interface. The computed curvature is constant everywhere except in the vicinity of the contact point. If we refine the surface representation near the contact line, we will observe an increase of the curvature—we noticed this behavior for all potentials applied.

The paper is organized as follows. In section 2, we present the electrowetting process and the plane capacitor approximation. In section 3, we derive our mathematical model. It is a shape inverse problem—we seek the drop shape such that it minimizes its total energy. The energy depends on the electric field, which is the solution of the external partial differential equation. The liquid volume is given and constant; it is considered an equality constraint. Finally, the problem consists in finding a min-max solution (saddle point) of an augmented Lagrangian (see [8]). Numerically, the solution is computed using Uzawa's algorithm and a quasi Newton optimization algorithm (BFGS). In section 4, we define the mathematical framework of shape optimization, and we derive the shape derivative of the augmented Lagrangian (the continuous gradient; see Theorem 4.1). In section 5, we detail the discretization of the equations and the shape derivative. The partial differential equation is solved using a standard linear P_1-Lagrange finite element method. The shape parameters and the shape deformation basis are defined; then the shape gradient and the optimization parameters are deduced from section 4. The full optimization process is presented in section 6. It has been implemented in C++. The code uses a public finite element library and a public mesh generator with automatic mesh refinement. In section 7, we present the algorithm we use to compute the droplet curvatures. It is based on a local least square approximation of the control points (second order Bezier approximation). In section 8, we present the numerical results.

2. Electrowetting process.

Let us consider the electrowetting process presented in Figure 2.1. We denote by σ_{LS}, σ_{SG}, and σ_{LG} the surface tension coefficients of the liquid-solid interface, solid-gas interface, and liquid-gas interface, respectively. We denote by θ the wetting angle.

When the applied electrical potential u_0 is null, Young's equation gives

$$\cos(\theta_0) = \frac{\sigma_{SG} - \sigma_{LS}}{\sigma_{LG}},$$

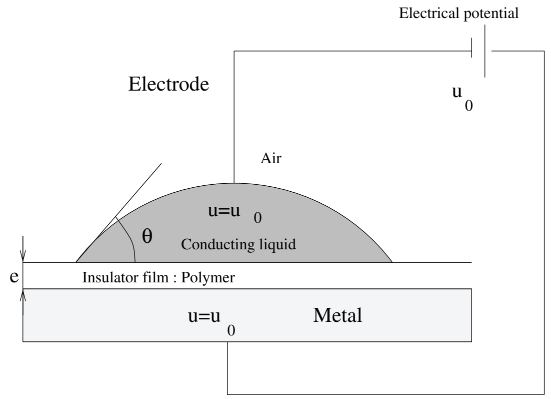


FIG. 2.1. *Electrowetting process.*

where θ_0 is the wetting angle at $u_0 = 0$.

Under the assumption that the system behaves as a plane capacitor with negligible boundary effects, the drop shape obeys the Young equation with the surface tension coefficient modified as follows (see [1]):

$$\sigma_{LS}(u_0) = \sigma_{LS} - \frac{\varepsilon_0 \varepsilon_1}{2e} u_0^2,$$

where e is the insulator thickness and ε_0 and ε_1 are the dielectric constants.

Also, we have (see [1])

$$\cos(\theta) = \cos(\theta_0) + \frac{\varepsilon_0 \varepsilon_1}{2\sigma_{LG}\ e} u_0^2.$$

This last equation is also called Lippman's equation.

Let us note that this law predicts total spreading when the potential increases. However, if u_0 is greater than a critical value u_{cr}, physicists observe a locking phenomenon limiting the spreading of the droplet on the polymer film. Such experiments are studied in [1], [21], [2].

The aim of the present study is to model and numerically compute the liquid drop shape for u_0 lower than the critical value u_{cr}. These computations include the wetting angle θ and the curvature κ of the liquid surface.

3. Mathematical modeling.

We model the electrowetting process described in the previous section as a shape inverse problem.

Assumptions.

(i) The applied electrical potential u_0 is continuous.

(ii) The liquid drop is a perfect conductor.

(iii) The drop geometry is 2D axisymmetric.

(iv) Electrostatic effects are negligible far away from the drop.

(v) For $u_0 = 0$, the liquid partially wets the polymer (the spreading coefficient is negative).

Notation (see Figure 3.1). We denote by $u(x)$ the electrical potential at point x, ω_0 the liquid drop, ω_1 the polymer domain, ω_2 the artificially bounded gas domain, and γ_{ext} its external boundary. The external boundary γ_{ext} is supposed to be located far enough from the liquid drop.

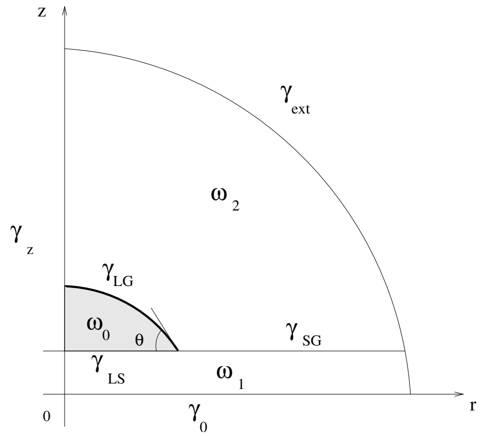


FIG. 3.1. *2D axisymmetric droplet (shaded gray). Domains and boundaries notations.*

We denote by γ_{LS}, γ_{SG}, and γ_{LG} the liquid-solid interface, solid-gas interface, and liquid-gas interface, respectively. We set $\omega = \omega_1 \cup \omega_2 \cup \gamma_{SG}$. We have $\partial\omega_0 = \gamma_{Lz} \cup \gamma_{LG} \cup \gamma_{LS}$ and $\partial\omega = \gamma_0 \cup \gamma_{Sz} \cup \gamma_{LG} \cup \gamma_{Gz} \cup \gamma_{ext}$, with $\gamma_z = \gamma_{Gz} \cup \gamma_{Lz} \cup \gamma_{Sz}$. We set $B = \omega_0 \cup \omega \cup \gamma_{LG} \cup \gamma_{LS}$. The liquid domain ω_0 will be variable; on the other hand, the domain B is given and fixed.

The questions we will answer numerically are the following. Given the electrical potential u_0,

- What is the drop shape?
- What is the wetting angle value θ?
- What is the curvature κ value of the drop surface?

The shape inverse formulation. We model this steady-state free surface problem as a shape inverse problem. We follow the approach presented in [4].

The total energy $\mathcal{E}$ is the sum of the gravitational energy, the capillary energy, and the electrostatic energy. In the 3D case, its expression is the following (see, e.g., [2]):

$$\mathcal{E}_{\omega_0} = \mathcal{E}_{\omega_0}^{grav} + \mathcal{E}_{\omega_0}^{cap} + \mathcal{E}_{\omega_0}^{elec};$$

with the gravitational energy:

$$\mathcal{E}^{grav} = \rho\ g \int_{\omega} z dx;$$

with the capillary balance energy:

$$\mathcal{E}^{cap} = \int_{\gamma_{LS}} (\sigma_{LS} - \sigma_{GS}) ds + \int_{\gamma_{LG}} \sigma_{LG} ds;$$

and the electrostatic energy:

$$\mathcal{E}^{elec} = -\frac{1}{2} \int_{\omega} \varepsilon |\nabla u|^2 dx,$$

where ρ is the liquid density, g is the gravity constant, $\varepsilon = \varepsilon_i$ in ω_i, $i = 1, 2$, and ε_i is the relative dielectric permittivity of ω_i; i.e., $\varepsilon_0\varepsilon_i$, $i = 1, 2$ is the polymer and the gas permittivity, respectively.

The shape inverse formulation is as follows:

$$\left\{\begin{array}{l} \text{Find } \omega_0^\star \text{ such that} \\ \mathcal{E}_{\omega_0^\star} \;=\; \displaystyle\min_{(\omega_0;\ \int_{\omega_0} dx = vol)} \mathcal{E}_{\omega_0}, \end{array}\right.$$

where *vol* is the given drop volume.

We set $u_i = u|_{\omega_i}$, $i = 1, 2$. Then, the potential u_i is the solution of the equation

$$-\mathrm{div}(\varepsilon_i \nabla u_i) \;=\; 0 \quad \text{in } \omega_i,\ i = 1, 2, \tag{3.1}$$

with the following Dirichlet boundary conditions:

$$\left\{\begin{array}{lcll} u_1 & = & u_0 & \text{on } \gamma_{LG}, \\ u_2 & = & u_0 & \text{on } \gamma_{LS}, \\ u_2 & = & 0 & \text{on } \gamma_0. \end{array}\right. \tag{3.2}$$

On the solid-gas interface, we have the transmission boundary conditions

$$\left\{\begin{array}{ccll} u_1 & = & u_2 & \text{on } \gamma_{SG}, \\ \varepsilon_1 \nabla u_1 n_1 & = & -\varepsilon_2 \nabla u_2 n_2 & \text{on } \gamma_{SG}. \end{array}\right. \tag{3.3}$$

On the artificial boundary $\gamma_{ext} = \gamma_{ext}^1 \cup \gamma_{ext}^2$, we impose

$$\varepsilon_i \nabla u_i n_i \;=\; 0 \qquad \text{on } \gamma_{ext}^i,\ i = 1, 2. \tag{3.4}$$

Therefore, the present mathematical problem is a shape optimal control problem for a system governed by a linear steady-state partial differential equation.

2*D axisymmetric equations.* As mentioned previously, we assume that the drop shape is 2D axisymmetric. We present below the weak formulation of the model. We set

$$X_0(\omega) = \{v \in H^1(\omega);\ v = 0 \text{ on } \gamma_0 \cup \gamma_{LS} \cup \gamma_{LG}\},$$

$$X_t(\omega) = \{v \in H^1(\omega);\ v = 0 \text{ on } \gamma_0;\ v = u_0 \text{ on } \gamma_{LS} \cup \gamma_{LG}\}.$$

The weak formulation of (3.1)–(3.4) in the 2D axisymmetric case is

$$\left\{\begin{array}{l} \text{Find } u^\omega \in X_t(\omega) \text{ such that} \\ \forall v \in X_0(\omega),\ a_\omega(u^\omega, v) = 0, \end{array}\right. \tag{3.5}$$

where

$$a_\omega(u, v) = \int_\omega \varepsilon r \langle \nabla u, \nabla v \rangle dx,$$

$x = (r, z)$, and $\langle .,. \rangle$ is the inner product of $\mathbb{R}^2$.

It follows from the Lax–Milgram theorem that state equation (3.5) has only one solution for $u^\omega \in X_t(\omega)$.

The shape inverse problem. In its dimensionless form, the drop energy is

$$\mathcal{E}_{\omega_0}(u^\omega) \;=\; \alpha \int_\omega z dx + \int_{\gamma_{LG}} r ds + \mu \int_{\gamma_{LS}} r dr - \delta \int_\omega \varepsilon r |\nabla u^\omega|^2 dx, \tag{3.6}$$

where u^ω is the unique solution of (3.5), $\alpha = \frac{\rho g (L^*)^2}{\sigma_{LG}}$, $\mu = -cos(\theta_0) = \frac{\sigma_{LS}-\sigma_{GS}}{\sigma_{LG}}$, $\delta = \frac{1}{2\sigma_{LG}L^*}$, and L^* is a characteristic length (typically $L^* \approx 10^{-4} - 10^{-3}$ m).

We set the cost function by

$$j(\omega) = \mathcal{E}_{\omega_0}(u^\omega). \tag{3.7}$$

We denote by $\mathcal{D}$ the admissible domain space. (The definition of $\mathcal{D}$ is detailed in the next section.) The shape optimal inverse problem is

$$\left\{ \begin{array}{l} \text{Find } \omega^\star \in \mathcal{D} \text{ such that} \\ j(\omega^\star) = \displaystyle\min_{(\omega; \int_{\omega_0} r dx = vol/2\pi)} j(\omega). \end{array} \right. \tag{3.8}$$

Let us point out that the variable is not the whole domain ω, but more precisely, the liquid-gas interface γ_{LG}; see Figure 3.1. We assume that the inverse shape problem (3.8) admits at least one solution. The existence of an optimal shape is not addressed in the present paper.

The augmented Lagrangian. Problem (3.8) is an optimization problem under an equality constraint. Thus, classically, we introduce the augmented Lagrangian $L_\tau : \mathcal{D} \times \mathbb{R} \longrightarrow \mathbb{R}$, defined by the following (see, e.g., [8]):

$$L_\tau(\omega, \lambda) = j(\omega) + \lambda c(\omega) + \tau c(\omega)^2, \tag{3.9}$$

where $c(\omega)$ is the volume constraint,

$$c(\omega) = \int_{\omega_0} r dx - \frac{vol}{2\pi} = \int_B r dx - \int_\omega r dx - \frac{vol}{2\pi}, \tag{3.10}$$

λ is the Lagrange multiplier, and τ is a penalty parameter.

Then, the shape optimal inverse problem (3.8) is formulated as the saddle-point problem:

$$\left\{ \begin{array}{l} \text{Find } (\omega^\star, \lambda^\star) \in \mathcal{D} \times \mathbb{R} \text{ such that} \\ L_\tau(\omega^\star, \lambda^\star) = \displaystyle\min_\omega \max_\lambda L_\tau(\omega, \lambda). \end{array} \right. \tag{3.11}$$

We will solve (3.11) using the classical Uzawa algorithm; see, e.g., [8]. This algorithm uses a gradient-type algorithm (BFGS), which requires us to compute the shape derivative of the cost function $\frac{dj}{d\omega}(\Omega)$ and the shape derivative of the constraint $\frac{dc}{d\omega}(\Omega)$. The expressions of these shape derivatives are presented in the next section.

4. Shape derivatives. As mentioned above, we need to compute the shape derivative of the cost function $\frac{dj}{d\omega}(\Omega)$ and the shape derivative of the constraint $\frac{dc}{d\omega}(\Omega)$. This is done using the optimal shape design method (see [15], [6], [11]; definitions of [7], [12] are used). Three approaches are possible: (i) we differentiate the equations and then we discretize them, thus obtaining the discretized continuous gradient; (ii) we discretize the equations and then we differentiate them, thus obtaining the discrete gradient; (iii) we directly differentiate the direct code (typically, using automatic differentiation). In the present study, we follow approach (i). This requires some extra mathematical definitions and tools, but this approach is rigorous; it leads to synthetic expressions of derivatives and it allows us to prove all the differentiabilities required. These derivatives are discretized in the next section, leading to shape gradients. The family of shapes considered is large enough in the sense that it includes those observed in the experiments.

This section is organized as follows. We define the admissible domain space $\mathcal{D}$ (Lipschitz domains); then we use the classical definition of shape derivatives based on domain deformations (a method of transport with $\mathcal{C}^1$ transformations). We prove the differentiability of the cost function j and the constraint function c with respect to the domain ω. Then, by introducing the adjoint state equation (in our case the adjoint state vanishes), we obtain the differential of j and c (Theorem 4.1). The shape derivative of the augmented Lagrangian L_τ follows (Corollary 4.2).

4.1. Mathematical framework: Domain variations and shape derivatives. We consider a family of Lipschitz domains. We define the space of admissible domains and the derivative with respect to the domain in a classical manner. The domain space is the set of domains homeomorphic to a reference domain. The transformations are $\mathcal{C}^1$ homeomorphisms. This regularity is necessary for all transported integrals to be well defined. The shape derivative of a real valued function is the derivative of the transported function with respect to the transformation. We refer to [15], [6], and we follow the definitions and properties presented in [7], [12].

Admissible domain space. Let $\hat{\Omega}$, a bounded open subset of $\mathbb{R}^2$ with a Lipschitz boundary, be the reference domain $\hat{\Omega} = \Omega_1 \cup \hat{\Omega}_2 \cup \hat{\Gamma}_{SG}$. Ω_1 represents the solid part and $\hat{\Omega}_2$ the gas part. We distinguish the variable part of $\hat{\Omega}$ from its fixed part; see Figure 4.1. We set $\partial\hat{\Omega} = \hat{\Gamma}_{Var} \cup \Gamma_{Fix}$, where $\hat{\Gamma}_{Var} = \hat{\Gamma}_{LG} \cup \hat{\Gamma}_{LS}$ is the variable boundary and Γ_{Fix} is the fixed boundary. We denote by B_{int} a neighborhood of $\hat{\Gamma}_{Var}$, B_{int} large enough; see Figure 4.1.

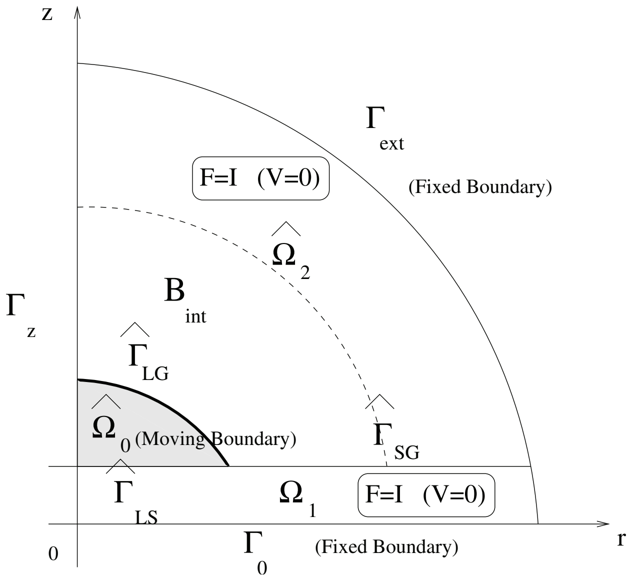


FIG. 4.1. *The reference domain $\hat{\Omega}$.*

We set the function space

$$\hat{\mathcal{F}} = \{\hat{F},\ \hat{F} \text{ bijection of } \hat{\Omega} \text{ onto } \hat{F}(\hat{\Omega});\ \hat{F} \in \mathcal{C}^1(\bar{\hat{\Omega}}, \mathbb{R}^d),\ \hat{F}^{-1} \in \mathcal{C}^1(\bar{\hat{F}}(\hat{\Omega}), \mathbb{R}^d)\} \tag{4.1}$$

and its affine subspace $\hat{\mathcal{F}}_0 = \{\hat{F} \in \hat{\mathcal{F}}; \hat{F} = I \text{ in } \hat{\Omega} \setminus B_{int}\}$, where I denotes the identity of $\mathbb{R}^d$. Then, we define the admissible domain space $\mathcal{D}$ as follows:

$$\mathcal{D} = \{\omega = \hat{F}_0(\hat{\Omega});\ \hat{F}_0 \in \hat{\mathcal{F}}_0\}. \tag{4.2}$$

One knows that if $\hat{F}$ is close enough to I in $\hat{\mathcal{F}}_0$ ($(\hat{F} - I)$ small enough), then $\hat{F}(\hat{\Omega})$ is an open set of $\mathbb{R}^2$ with a Lipschitz boundary and $F(\hat{\Gamma}_{Var}) \subset B_{int}$.

Shape derivative of a real valued function. For $\hat{F}_0 \in \hat{\mathcal{F}}_0$, $(\hat{F}_0 - I)$ small enough, we define the domain Ω by $\Omega = \hat{F}_0(\hat{\Omega})$ and $\Gamma_{Var} = \hat{F}_0(\hat{\Gamma}_{Var})$. We set the homeomorphism space defined in Ω (see Figure 4.2) as $\mathcal{F} = \{F,\ F = \hat{F} \circ \hat{F}_0^{-1},\ \hat{F} \in \hat{\mathcal{F}}\}$, and its affine subspace as $\mathcal{F}_0 = \{F,\ F = \hat{F} \circ \hat{F}_0^{-1},\ \hat{F} \in \hat{\mathcal{F}}_0\}$.

Let $F \in \mathcal{F}_0$; we define $\omega = F(\Omega)$ and $V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^d)$ by $V = F - I$. We have $V = 0$ in $\hat{\Omega} \setminus B_{int}$.

$$\begin{tikzcd}
\hat{\Omega} \arrow[r, "\hat{F}_0"] \arrow[dr, "\hat{F}"'] & \Omega \arrow[d, "F= I+ V"] \\
& \omega
\end{tikzcd}
\qquad
\begin{tikzcd}
\Omega \arrow[rr, shift left, "F= I+ V"] \arrow[dr, "\bar{v}"'] & & \omega \arrow[ll, shift left, "F^{-1}"] \arrow[dl, "v"] \\
& R &
\end{tikzcd}$$

FIG. 4.2. *Change of variables.*

For a given cost function j, $j : \omega \in \mathcal{D} \mapsto j(\omega) \in \mathbb{R}$, we define the "transported" cost function $\bar{\jmath}$ by $\bar{\jmath} : \mathcal{F}_0 \to \mathbb{R}\ :\ F \mapsto \bar{\jmath}(F) = j(F(\Omega)) = j(\omega)$. Then, the derivative with respect to the domain is defined as follows (see, e.g., [15], [7] for more details):

$$\frac{dj}{d\omega}\ (\Omega) \cdot V = \frac{d\bar{\jmath}}{dF}\ (I) \cdot V \quad \forall V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^d). \tag{4.3}$$

4.2. Shape derivatives. We present below the expressions of the exact differentials with respect to the shape ω.

THEOREM 4.1. *There exists $\mathcal{V}_I$, a neighborhood of I in $\mathcal{F}_0$, such that*

(i) *the cost function $j : \mathcal{D} \to \mathbb{R}$; $\omega \mapsto j(\omega) = \mathcal{E}_{\omega_0}(u^\omega)$ belongs to $\mathcal{C}^1$ for all $\omega = F(\Omega)$, $F \in \mathcal{V}_I$. Additionally, for all $V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^2)$, we have*

$$\frac{dj}{d\omega}(\Omega).V = \frac{\partial \mathcal{E}_{\Omega_0}}{\partial \omega}(u^\Omega).V, \tag{4.4}$$

with u^Ω the solution of the state equation (3.5) *posed in Ω and*

$$\begin{aligned}
\frac{\partial \mathcal{E}_{\Omega_0}}{\partial \omega}(u^\Omega).V = {} & \alpha \int_\Omega z \circ V\ dx + \alpha \int_\Omega z \operatorname{div}(V)\ dx \\
& + \int_{\Gamma_{LG}} r \circ V\ ds + \int_{\Gamma_{LG}} r\ \operatorname{div}_\Gamma V\ ds \\
& + \mu \int_{\Gamma_{LS}} r \circ V\ dr + \mu \int_{\Gamma_{LS}} r\ \operatorname{div}_\Gamma V\ dr \\
& - \delta \int_\Omega \varepsilon\ (r \circ V)\ |\nabla u^\Omega|^2\ dx - \delta \int_\Omega \varepsilon r |\nabla u^\Omega|^2\ \operatorname{div}(V)\ dx \\
& + \delta \int_\Omega \varepsilon r < (\ {}^T DV + DV) \nabla u^\Omega, \nabla u^\Omega > dx,
\end{aligned}$$

with $\operatorname{div}_\Gamma V = (\operatorname{div}(V) - \langle n,\ {}^T DVn \rangle)$, *$n$ the external normal, and $x = (r, z)$.*

(ii) *The volume constraint $c(\omega)$ belongs to $\mathcal{C}^1$ for all $\omega = F(\Omega)$, $F \in \mathcal{V}_I$ and, for all $V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^2)$,*

$$\frac{dc}{d\omega}(\Omega).V = -\int_\Omega r \circ V dx - \int_\Omega r\mathrm{div}(V)dx. \tag{4.5}$$

Proof. The proof follows with three steps: 1. transport of equations; 2. differentiability with respect to ω; 3. use of the adjoint technique leading to the expression of the exact differential.

Step 1. *Transport of equations.* As noted previously, we need to transport the cost function j in order to compute its shape derivative. To this end, we need to transport all the equations on the reference domain $\Omega = F^{-1}(\omega)$.

For any $u, v \in X_0(\omega)$, we let

$$\bar{a}(F;\bar{u},\bar{v}) = a_{F(\Omega)}(\bar{u} \circ F^{-1}, \bar{v} \circ F^{-1}) = a_\omega(u,v)$$
$$= \int_\Omega \bar{\varepsilon}\bar{r} <^T (DF^{-1} \circ F)\nabla\bar{u},^T (DF^{-1} \circ F)\nabla\bar{v} > |\mathrm{det}DF| d\bar{x},$$

with $\bar{u} = u \circ F$, $\bar{v} = v \circ F$, $\bar{x} = x \circ F$, and $\bar{\varepsilon} = \varepsilon \circ F$; see Figure 4.2.

The mapping $v \in X_0(F(\Omega)) \mapsto v \circ F \in X_0(\Omega)$ is an isomorphism for $F \in \mathcal{F}_0$. In other respects, the Dirichlet data u_0 is constant; hence $u_0 = u_0 \circ F$. Then, since state equation (3.5) has a unique solution u^ω, the transported state equation

$$\text{Find } \bar{u}^F \in X_t(\Omega): \ \bar{a}(F;\bar{u},\bar{v}) = 0 \quad \forall \bar{v} \in X_0(\Omega)$$

has a unique solution $\bar{u}^F = u^\omega \circ F$.

Similarly, for any $u \in X_0(\omega)$, we let $\bar{\mathcal{E}}(F;\bar{u}) = \mathcal{E}_{F(\Omega_0)}(\bar{u} \circ F^{-1}) = \mathcal{E}_{\omega_0}(u)$. We have $\bar{\jmath}(F) = \bar{\mathcal{E}}(F;\bar{u}^F)$,

$$\begin{aligned} \bar{\jmath}(F) = \alpha &\int_\Omega \bar{z}\ |\mathrm{det}DF|\ d\bar{x} \\ &+ \int_{\Gamma_{LG}} \bar{r}\ Jac(F)\ d\bar{s} + \mu \int_{\Gamma_{LS}} \bar{r}\ Jac(F)\ d\bar{r} \\ &- \delta \int_\Omega \bar{\varepsilon}\ \bar{r}\ |\ ^T(DF^{-1} \circ F)\nabla\bar{u}^F|^2\ |\mathrm{det}DF|\ d\bar{x}, \end{aligned} \tag{4.6}$$

with $Jac(F) = |\mathrm{det}DF|\ \|\ ^T DF^{-1}.n\|_{\mathbb{R}^2}$.

Also, we define

$$\bar{c}(F) = \int_B r dx - \int_\Omega \bar{r}\ |\mathrm{det}DF|\ d\bar{x} - \frac{vol}{2\pi}. \tag{4.7}$$

Step 2. *Differentiability with respect to ω.* The mapping $\bar{a}(F;\bar{u},\bar{v})$ is C^1 with respect to $(F;\bar{u})$. It follows from the implicit function theorem that the transported state equation defines a C^1-mapping $F \mapsto \bar{u}^F : \mathcal{F}_0 \to X_t(\Omega)$ in a neighborhood $\mathcal{V}_I$ of I.

Then, since the mapping $\bar{\mathcal{E}}$ is of class $\mathcal{C}^1(\mathcal{F} \times X_0(\Omega))$, the cost function j is continuously differentiable. Also, the constraint function c is continuously differentiable.

Step 3. *Expression of the exact differential.* By definition, we have $\frac{dj}{d\omega}(\Omega) \cdot V = \frac{d\bar{\jmath}}{dF}(I) \cdot V$ for all $V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^2)$.

Then, using the classical adjoint technique, we have

$$\frac{d\bar{\jmath}}{dF}(I).V = \frac{\partial \bar{\mathcal{E}}}{\partial F}(I;u^{\Omega}).V - \frac{\partial \bar{a}}{\partial F}(I;u^{\Omega},p^{\Omega}).V \quad \forall V \in \mathcal{C}^1(\bar{\Omega},\mathbb{R}^2),$$

where u^{Ω} is the solution of the state equation posed in Ω and $p^{\Omega} \in X_0(\Omega)$ is the adjoint state, unique solution of the following adjoint equation:

$$\frac{\partial \bar{a}}{\partial u}(I;u^{\Omega},p^{\Omega}).v = \frac{\partial \bar{\mathcal{E}}}{\partial u}(I;u^{\Omega}).v \quad \forall v \in X_0(\Omega).$$

We have

$$\frac{\partial \bar{a}}{\partial u}(I;u^{\Omega},p^{\Omega}).v = a_{\Omega}(p^{\Omega},v) \quad \text{and} \quad \frac{\partial \bar{\mathcal{E}}}{\partial u}(I;u^{\Omega}).v = -2\delta a_{\Omega}(u^{\Omega},v) = 0 \ \ \forall v \in X_0(\Omega).$$

Hence, $p^{\Omega} \in X_0(\Omega)$ and $a_{\Omega}(p^{\Omega},v) = 0$ for all $v \in X_0(\Omega)$. Therefore, $p^{\Omega} = 0$.

Hence,

$$\frac{dj}{d\omega}(\Omega).V = \frac{\partial \bar{\mathcal{E}}}{\partial F}(I;u^{\Omega}).V \quad \forall V \in C^1(\bar{\Omega},\mathbb{R}^2).$$

Using (4.7) and the classical expression of the derivatives of $|\det(DF)|$, $(DF^{-1} \circ F)$, and $(\| {}^{T}DF^{-1}.n\|_{\mathbb{R}^2})$ (see, e.g., [15, Chap. IV]), we obtain the result (i).

The result (ii) follows from (4.7) and the expression of the derivative of $|\det(DF)|$. □

Then, we have straightforwardly the following result.

COROLLARY 4.2. *At (λ,τ) given in $\mathbb{R} \times \mathbb{R}^+$, the augmented Lagrangian L_τ is locally and continuously differentiable with respect to ω. And for all $V \in \mathcal{C}^1(\bar{\Omega},\mathbb{R}^2)$,*

$$\frac{\partial L_\tau}{\partial \omega}(\Omega,\lambda).V = \frac{dj}{d\omega}(\Omega).V + \lambda\frac{dc}{d\omega}(\Omega).V + 2\tau c(\Omega)\frac{dc}{d\omega}(\Omega).V, \tag{4.8}$$

where $\frac{dj}{d\omega}(\Omega).V$ and $\frac{dc}{d\omega}(\Omega).V$ are defined by (4.4) *and* (4.5)*, respectively.*

5. Discretization.

In this section, we discretize the shape derivative of the augmented Lagrangian L_τ defined by (4.8); then we define the shape parameters and obtain the shape gradient. Then, we detail the full optimization process. We follow [7], [12]; see also [13].

Let us recall that the expression $\frac{\partial L_\tau}{\partial \omega}(\Omega,\lambda).V$ depends on u, the unique solution of (3.5).

Let $(\mathcal{T}_h)$ be a regular family of triangulation, where $\omega = \cup_{T\in\mathcal{T}_h} T$. We compute an approximation of u using the classical piecewise linear conforming finite element method (P_1-Lagrange). This finite element approximation is denoted by u_h, where the parameter h denotes a characteristic mesh size.

Discretization of the boundary and the shape parameters. Let $\hat{\Omega}$ be an open set of reference; typically $\hat{\Omega}$ is a quarter of a disk; see Figure 4.1. The domain of reference $\hat{\Omega}$ is defined using a parametric function:

$$s_{\hat{\Omega}}(t) = \sum_{i=0}^{N-1} \hat{P}_i \, s_i(t)\,,\ t \in [0,1],$$

where $\{s_i(t)\}_{i=0,\ldots,N-1}$ are piecewise linear functions, $s_i(\frac{j}{N-1}) = \delta_{ij}$; δ_{ij} denotes the Kronecker symbol, and $\hat{P}_i = ((\hat{P}_r)_i,(\hat{P}_z)_i)^T$ are the control points. We set $(\hat{P}_z)_1 = (\hat{P}_z)_0$.

We have $\Omega = \hat{F}_0(\hat{\Omega})$ with $\hat{F}_0 \in \hat{\mathcal{F}}_0$. Similarly, we define the variable boundary Γ_{LG} (the unknown of the problem) by

$$s_\Omega(t) = \sum_{i=0}^{N-1} P_i\, s_i(t)\,,\ t \in [0,1].$$

Hence, the boundary Γ_{LG} is defined by N control points $P_i, i = 0, \ldots, N-1$.

Initially, these points define $\hat{\Gamma}_{LG}$ as follows (see Figure 5.1):

$$\hat{P}_i = (0, R)^T,$$

$$\hat{P}_i = \left(R \cos\left(\frac{(N-1-i)\pi}{2(N-1)}\right),\ R \sin\left(\frac{(N-1-i)\pi}{2(N-1)}\right)\right)^T, \quad i = 2, \ldots, N-1,$$

$$\hat{P}_1 = \left(R \cos\left(\frac{(N-2)\pi}{2(N-1)}\right)/2, R\right)^T.$$

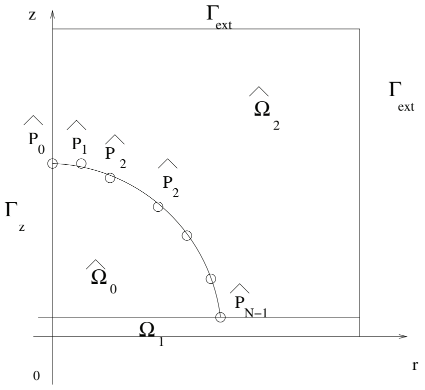


FIG. 5.1. *Reference domain. Parametrization.*

Therefore, during the optimization process, we compute a new domain that requires computing new control points P_i, $i = 0, \ldots, N-1$.

The shape deformation space. Let us discretize the shape deformation V, $V \in \mathcal{C}^1(\bar{\Omega}, \mathbb{R}^2)$. We have $\Omega = \hat{F}_0(\hat{\Omega})$ with $\hat{F}_0 \in \hat{\mathcal{F}}_0$. We set $V = \hat{V} \circ \hat{F}_0^{-1}$. V is defined in Ω, while $\hat{V}$ is defined in $\hat{\Omega}$.

We approximate $C^1(\bar{\hat{\Omega}}, \mathbb{R}^2)$ by $\hat{S}_H$, the vectorial space spanned by $\{\hat{V}_i\}_{i=0,\ldots,N-1}$:

$$\hat{S}_H = Span\{\hat{V}_i\}_{i=0,\ldots,N-1},$$

where the set of vectors $\{\hat{V}_i\}_{i=0,\ldots,N-1}$ is detailed below.

We set $H = \frac{1}{N-1}$. The parameter H denotes a characteristic size of the shape deformation space.

Then, the deformation field V is approximated by

$$V_H = \sum_{i=0}^{N-1} \eta_i V_i, \tag{5.1}$$

where $V_i = \hat{V}_i \circ \hat{F}_0^{-1}$ and η_i, $i = 0, \ldots, N-1$ are real coefficients.

We have $V_H = (\hat{V}_H \circ \hat{F}_0^{-1})$ with

$$\hat{V}_H = \sum_{i=0}^{N-1} \eta_i \hat{V}_i. \tag{5.2}$$

Finally, $C^1(\bar{\Omega}, \mathbb{R}^2)$ is approximated by $S_H = Span\{V_i = \hat{V}_i \circ \hat{F}_0^{-1}\}_{i=0,\dots,N-1}$.

The shape deformation basis. We have $\hat{F}_0 = (I + \hat{V})$, and $\hat{V}$ is approximated by $\hat{V}_H$, which was defined by (5.2).

The basis $\{\hat{V}_i\}_{i=0,\dots,N-1}$, is defined in $\hat{\Omega}$ as follows. For $i = 0, \dots, N-1$, we solve

$$\left\{ \begin{array}{ccll} \Delta(\hat{V}_r)_i & = & 0 & \text{in } \hat{\Omega} \cap Bint, \\ (\hat{V}_r)_i & = & 0 & \text{in } \hat{\Omega}/Bint, \\ (\hat{V}_r)_i & = & 0 & \text{on } \Gamma_{Gz} \cup \Gamma_{Sz}, \\ (\hat{V}_r)_i & = & \frac{(\hat{P}_r)_i}{||\hat{P}_i||}\, s_i & \text{on } \hat{\Gamma}_{LG}, \end{array} \right. \tag{5.3}$$

$$\left\{ \begin{array}{ccll} \Delta(\hat{V}_z)_i & = & 0 & \text{in } \hat{\Omega} \cap Bint, \\ (\hat{V}_z)_i & = & 0 & \text{in } \hat{\Omega}/Bint, \\ (\hat{V}_z)_i & = & 0 & \text{on } \Gamma_0 \cup \hat{\Gamma}_{LS} \cup \hat{\Gamma}_{SG}, \\ (\hat{V}_z)_i & = & \frac{(\hat{P}_z)_i}{||\hat{P}_i||}\, s_i & \text{on } \hat{\Gamma}_{LG}, \end{array} \right. \tag{5.4}$$

where $\hat{V}_i = ((\hat{V}_r)_i, (\hat{V}_z)_i)^T$, $\hat{P}_i = ((\hat{P}_r)_i, (\hat{P}_z)_i)^T$, and $\|\hat{P}_i\| = [(\hat{P}_r)_i^2 + (\hat{P}_z)_i^2]^{\frac{1}{2}}$.

Let us note that we could have extended this vector field over the whole domain by solving a linear elasticity system.

The shape gradient. We approximate V by V_H (see (5.1)), and we have

$$\frac{\partial L_\tau}{\partial \omega}(\Omega, \lambda).V \approx \frac{\partial L_\tau}{\partial \omega}(\Omega, \lambda).V_H = \sum_{i=0}^{N-1} \eta_i \frac{\partial L_\tau}{\partial \omega}(\Omega, \lambda).V_i.$$

Then, the shape gradient denoted by G^H is the vector

$$\begin{aligned} G^H = (G_i^H)_{i=0,\dots,N-1} &= \left(\left[\frac{\partial L_\tau}{\partial \omega}(\Omega, \lambda).V_i \right] \right)_{i=0,\dots,N-1} \\ &= \left(\left[\frac{\partial L_\tau}{\partial \omega}(\Omega, \lambda).(\hat{V}_i \circ \hat{F}_0^{-1}) \right] \right)_{i=0,\dots,N-1}, \end{aligned}$$

where $\Omega = \hat{F}_0(\hat{\Omega})$.

Finally, we have for all $i = 0, \dots, N-1$ (see Corollary 4.2),

$$G_i^H = \frac{dj}{d\omega}(\Omega).(\hat{V}_i \circ \hat{F}_0^{-1}) + \lambda \frac{dc}{d\omega}(\Omega).(\hat{V}_i \circ \hat{F}_0^{-1}) + 2\tau c(\Omega) \frac{dc}{d\omega}(\Omega).(\hat{V}_i \circ \hat{F}_0^{-1}). \tag{5.5}$$

Variables of optimization. Since $\Omega = \hat{F}_0(\hat{\Omega}) = (I + \hat{V})(\hat{\Omega}) \approx (I + \hat{V}_H)(\hat{\Omega})$ with $\hat{V}_H$ defined by (5.2), and $\hat{V}_i$ defined by (5.3), (5.4), the variables of optimization are the N coefficients η_i, $i = 0, \dots, N-1$.

6. Optimization process. As mentioned previously, we solve (3.11), an optimization problem with constraint, using Uzawa's algorithm; see, e.g., [8]. This algorithm requires a descent algorithm which is in the present case BFGS (the quasi Newton method). This gives the following:

- Initially, we set $\eta_i^0 = 0$, $i = 0, \dots, N-1$; $\lambda_0 = 0$.
- We compute the volume constraint $c(\eta^0)$.
- While the volume constraint ($|c(\eta^{k+1})| > eps1$) is not satisfied,
  - set $\lambda_{k+1} = \lambda_k + \rho\, c(\eta^k)$,
  - compute η_i^{k+1} $i = 0, \dots, N-1$ such that $L_\tau(\eta^{k+1}, \lambda_{k+1}) < L_\tau(\eta^k, \lambda_{k+1})$ using the BFGS algorithm, and
  - compute the volume constraint $c(\eta^{k+1})$.

Classically, we set $\rho = \tau$; see [8].

The BFGS algorithm is implemented with bounding constraints. The linear search is done using a dichotomic process.

We stop the BFGS algorithm either if $\frac{|j(\eta^{k+2}) - j(\eta^{k+1})|}{j(\eta^{k+1})} < eps2$ or if $\|(G^H)^{k+2}\| < eps3$.

As usual, each call of the algorithm BFGS implies a few calls to the simulator.

The simulator does the following:

- It computes the new shape and the new mesh defined by

$$\Omega = \left(I + \sum_{i=0}^{N-1} \eta_i \hat{V}_i \right) (\hat{\Omega}).$$

- It solves the state equation (3.5) posed in Ω by a P_1-Lagrange finite element method (with or without automatic mesh refinement).
- It computes the augmented Lagrangian L_τ defined by (3.9), with its gradient G^H defined by (5.5), and the volume constraint c defined by (3.10).

The full optimization process is represented in Figure 6.1.

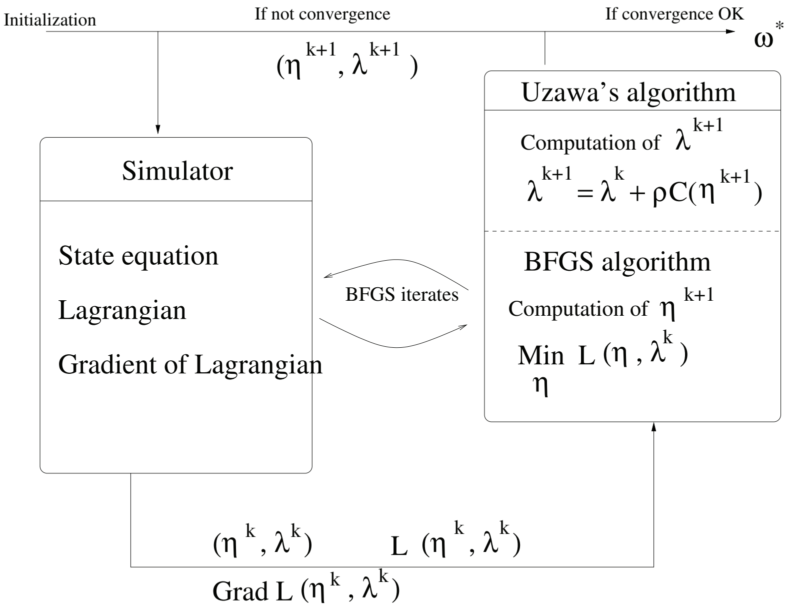


FIG. 6.1. *The optimization process.*

7. Curvature computation. In the next section, we consider the evaluation of the droplet curvature, particularly near the contact line. It was shown in [5] that the contact angle approaches Young's angle, independently of the applied electrical potential value. Observations show that the curvature is not constant. Then, it would be interesting to see if the present modeling approach allows us to observe such changes of curvature values near the triple point.

Accurately computing the droplet curvature is a difficult task since its interface is a piecewise linear curve, and hence is not C^2 differentiable. In addition, points defining this piecewise linear curve result from the full shape optimal design process and hence may comprise some nonnegligible numerical errors. Thus, we seek to estimate the curvature of an underlying smooth surface.

Computing a discrete surface curvature is a classical (and difficult) problem. Usually in the computer aided geometric design context, surfaces are 3D and triangularized, and the objectives are to smooth the mesh and simplify it, but not to quantify a local variation of curvature; see, e.g., [10].

We are facing the following dilemma. We seek to get rid of numerical errors on the points defining the curve while we try to detect as accurately as possible a local significant variation of curvature.

We do not consider a direct computation by a finite difference method since it is very sensitive to data error. We do not consider a polynomial reconstruction of the underlying smooth surface and then evaluate its curvature, since this leads to inaccurate results and unexpected behavior. Following [9], [14], we consider a local least square approximation and then we evaluate the curvature. In the present algorithm, we consider a second order local Bezier approximation; see [14]. As the numerical tests presented below show, this method filters noise reasonably.

7.1. The algorithm. Given N points $X_i = (r_i, z_i)^T$, $i = 1, \ldots, N$ defining the liquid-gas interface, the basic idea is to approximate these data using a local least square approximation by a Bezier curve.

The Bezier curve $\mathcal{C}(t)$ is given by

$$\mathcal{C}(t) = (r(t), z(t))^T = \sum_{j=1}^{M} P_j B_{j-1}^{M-1}(t) \qquad \text{for } t \in [0,1],$$

where $P_j = (\alpha_j, \beta_j)^T \in \mathbb{R}^2$ are the control points and $\{B_j^m(t)\}_{0 \le j \le M-1}$ is the classical Bernstein basis, with $B_j^m \in P_m$, $B_j^m(t) = C_j^m (1-t)^{m-j} t^j$, C_j^m being the binomial coefficients.

We set $M = 3$; hence we consider three points of control P_j and second degree curves.

For an inner point X_i (see Figure 7.1), we compute the least square approximation of the four points $\{X_{i-2}, \ldots, X_{i+2}\}$ by Bezier's curve as follows. We minimize

$$J(P_1, P_2, P_3) = \sum_{l=i-2}^{i+2} \left\| \sum_{j=1}^{3} P_j B_{j-1}^2(t_l) - X_l \right\|^2,$$

where $\{t_{i-2} = 0, \ldots, t_{i+2} = 1\}$ is a uniform subdivision of $[0,1]$. The unique minimum is computed by solving the corresponding normal equations.

For the extremal point X_1, we consider a Bezier curve approximating the points X_i for $i = 1, \ldots, 4$. For X_2, we consider a Bezier curve approximating the points X_i for $i = 1, \ldots, 5$.

For the extremal points X_{N-1} and X_N, the principle is similar.

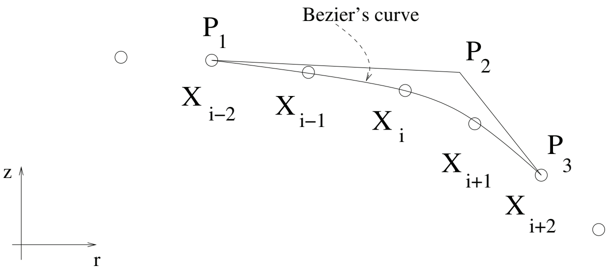


FIG. 7.1. *Inner point X_i. Local least square approximation using Bezier's curve.*

Curvature expression. Once a Bezier curve $\mathcal{C}(t) = (r(t), z(t))^T$ is computed for each point X_i, we evaluate the curvature as follows:

$$\kappa_i \;\equiv\; \kappa(t_i) = \frac{r'z" - r"z'}{(r'^2 + z'^2)^{\frac{3}{2}}}(t_i),$$

where (r', z') and (r'', z'') are computed using de Casteljau's algorithm, with t_i being the parameter value related to X_i.

Sensitivity to random noise. Since the control points defining the (optimal) droplet shape result from the full optimization process, they are perturbed by some nonnegligible numerical errors. Hence, we test the robustness of our algorithm to data perturbation below.

We set $N(r, z) = (r'z" - r"z')$ and $D(r, z) = (r'^2 + z'^2)^{\frac{3}{2}}$; hence $\kappa(r, z)(t) = \frac{N(r,z)}{D(r,z)}(t)$. Let δz be a perturbation on the z-coordinate of data X_i, $i = 1, \ldots, N$; then we have

$$\frac{\partial \kappa}{\partial z}(r, z).\delta z = \frac{N(r, \delta z)}{D(r, z)} - 3\ \frac{\kappa(r, z)}{(r'^2 + z'^2)} z'\delta z.$$

This formula expresses the curvature sensitivity to perturbation on z-coordinates. Noise introduced below is a random perturbation on the z-coordinate of data X_i, $i = 1, \ldots, N$. It is a normal distribution with mean zero and variance one.

7.2. Numerical tests. The numerical tests presented below are useful for (i) validating the present algorithm on explicit curves knowing their curvature value (the "exact" curves); and (ii) measuring the computed curvature sensitivity to random perturbation on data.

To this end, we consider an "oscillating curve" (see Figure 7.2), defined by N points as follows:

$$r(s) = (R + \epsilon\cos(a.s))\cos\left(\frac{\pi}{2}s\right), \quad z(s) = (R + \epsilon\cos(a.s))\sin\left(\frac{\pi}{2}s\right),$$

with $s \in [0, 1]$, s discretized by N points similarly to η and $\epsilon = \frac{R}{10}$, $a = 10$, $R = 1$.

The exact curvature of the "oscillating circle" is straightforwardly obtained. This curve presents smooth variations of curvature with changes of sign. If we compare the curvature values computed by the present algorithm and those computed by the second order finite difference scheme directly applied to the N data $X_i = (r_i, z_i)^T$, $i = 1, \ldots, N$, then without noise both lead to similar accuracy—the two methods give a very precise approximation.

However, in the presence of noise, the direct approximation does not give any good approximation. On the contrary, the present algorithm, based on a local least square approximation of the surface by Bezier's curve, leads to a good approximation of the curvature value of the nonperturbed curve.

We present in Figure 7.2 the curvature values obtained with the present algorithm when some noise is introduced. As mentioned above, the noise is defined as a perturbation on the z-coordinate of data X_i, $i = 1, \dots, N$. Its magnitude is about 0.5%.

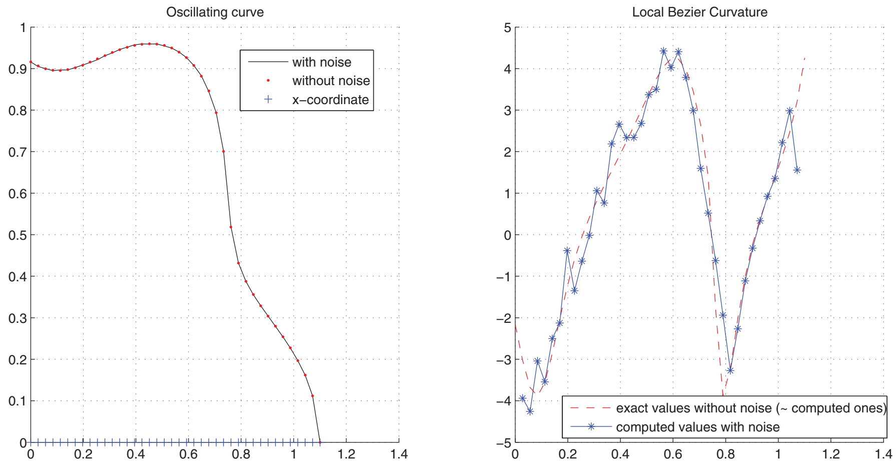


FIG. 7.2. *Left: Oscillating curve. Right: Computed curvature value when noise is introduced.*

8. Numerical results.

The full optimization process described in the previous section was implemented in C++. Our software, *ElectroCap* (see [13]), is based on the public C++ finite element library Rheolef [19] and an in-house BFGS algorithm. The mesh generator used is Bamg. For each simulator call, an automatic mesh refinement is used. This mesh refinement is based on the classical a posteriori estimates. We present in Figure 8.2 a typical mesh with the adaptive mesh in the edge.

Numerical data. The numerical data considered are the following:

- the surface tension coefficients (in N/m): $\sigma_{LS} = 2.7\ 10^{-2}$, $\sigma_{LG} = 5\ 10^{-2}$;
- the wetting angle at $u_0 = 0$ (in radians): $\theta_0 = \frac{\pi}{2}$ (hence $\mu = 0$);
- the insulator thickness (in m): $e = 200\ 10^{-6}$;
- the electrical permitivities: $\varepsilon_1 = 2 \times 8.85\ 10^{-12}$ and $\varepsilon_2 = 8.85\ 10^{-12}$;
- the drop volume (in L): $vol = 40\ 10^{-9}$.

We assume that the Bond number α is small; i.e., we neglect the gravitational term. Then, the cost function is (see (3.6), (3.7))

$$j(\omega) = j_{cap}(\omega) - j_{elec}(\omega), \tag{8.1}$$

with

$$j_{cap}(\omega) = \int_{\gamma_{LG}} r ds \quad \text{and} \quad j_{elec}(\omega) = \delta \int_{\omega} \varepsilon r |\nabla u^{\omega}|^2 \, dx,$$

where $j_{cap}(\omega)$ and $j_{elec}(\omega)$ are positive cost functions. The numerical parameters are the following:

- the penalty parameter: $\tau = \rho = 10^{-3}$;
- the convergence parameter of Uzawa's algorithm: $eps1 = 10^{-3}$;
- the convergence parameter of the BFGS algorithm: $eps2 = eps3 = 10^{-3}$;
- the number of control points: $NCP = 50$.

The NCP is defined as follows. If we consider the polar coordinates in the plane, for a droplet of radius R, the N points are equidistributed in θ. Their positions are indicated in Figure 8.4.

Code validation. All components of the code have been validated—the direct problem, the transport of the mesh, and the shape gradient.

The computed shape gradient has been compared with values obtained by a finite difference method using the following approach. For each shape parameter, a finite difference shape derivative is computed using a domain perturbation of magnitude 10^{-4}. The order of magnitude of the relative error obtained between the two approaches is between 10^{-4} and 10^{-6}, depending on the imposed electrical field value u_0.

In order to validate the entire code, we simulate the Lippman approximation by using the code with $u_0 = 0$ V but changing σ_{LS} for each value of u_0 using the formula given by the approximation of the plane capacitor:

$$\sigma_{LS}(u_0) = \sigma_{LS} - \frac{\varepsilon_0 \varepsilon_1}{2e} u_0^2.$$

Thus, theoretically, the contact angle should also be given by the Lippman equation. Numerically we observe a good agreement with the theory. Figure 8.1 shows the value of the contact angle found numerically (the angle of the last mesh triangle) and the theoretical value given by the Lippman equation.

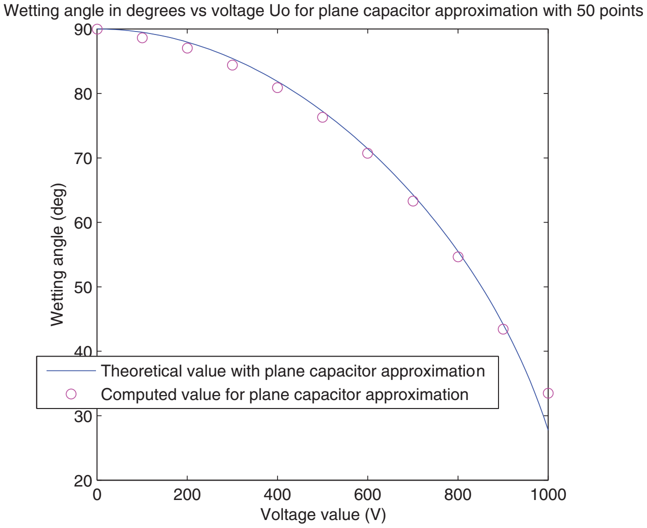


FIG. 8.1. *Plane capacitor approximation.*

Moreover, we compute the curvature for each value of σ_{LS} (which corresponds to a value of u_0). Given a value of u_0 and thus a value of σ_{LS}, we notice that the numerically computed curvature remains constant for each point of the drop. We also obtain for this case a very good agreement with the theory, which contributes to validating the code.

Now we compute the drop shape with the initial model, i.e., by considering σ_{LS} as a constant and by changing values of u_0.

Drop shape and wetting angle. We present in Figure 8.2 the drop shape (with mesh) obtained for $u_0 = 400$ V (left) and a zoom of the refined mesh near the edge (right). As a matter of fact, we use an adaptive mesh refinement near the contact point based on a posteriori estimates. All meshes contain approximately 4000 elements and 2000 vertices. For each computation, the volume constraint is satisfied at less than 0.1%.

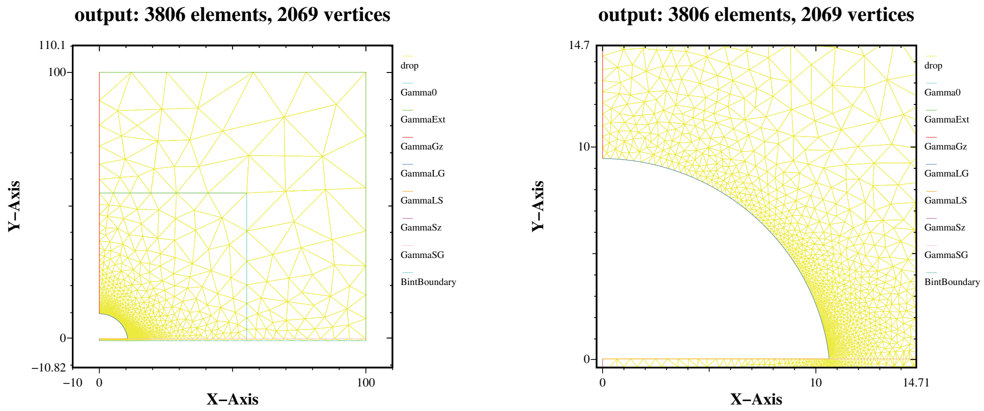


FIG. 8.2. *Left: Shape and mesh for* $u_0 = 400$ *V. Right: Zoom near the drop.*

We present the cost function, the augmented Lagrangian, and its gradient as a function of the iteration number for $u_0 = 400$ V in Figure 8.3. The behavior of the algorithm for other values of u_0 is similar.

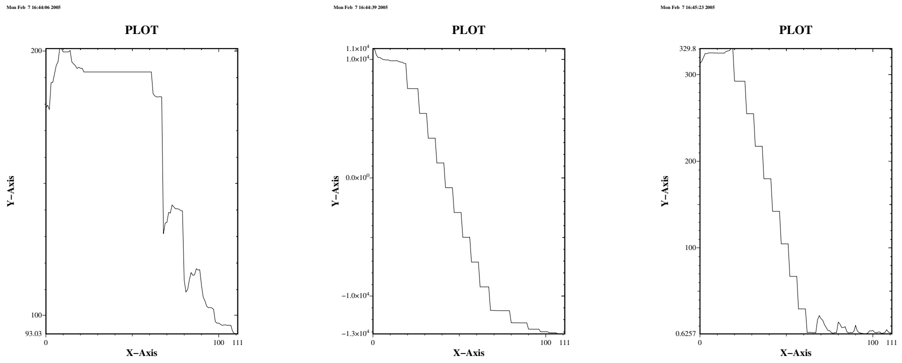


FIG. 8.3. $u_0 = 400$ *V. Left. Cost function* j *versus iterations. Middle. Augmented Lagrangian* L_τ*. Right. Gradient of* L_τ*.*

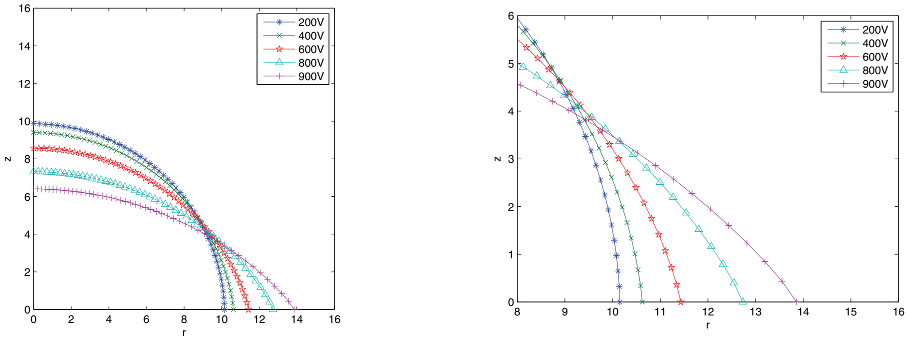


FIG. 8.4. *Droplet surfaces for different u_0 values. At right: zoom near the triple point.*

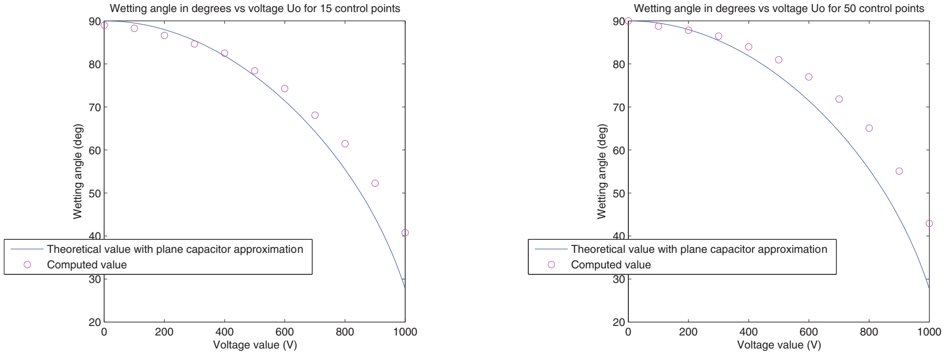


FIG. 8.5. *Wetting angle. Computed values and Lippman's equation predictions. Left. With $NCP = 15$. Right. With $NCP = 50$.*

We present in Figure 8.4 the drop shapes obtained in the function of u_0.

We present in Figure 8.5 (left and right) the wetting angle values in the function of u_0. In both figures (left and right), we plot the computed values (the angle of the last mesh triangle) and values predicted by the Lippman equation. On the left, plotted values are obtained using 15 control points ($NCP = 15$); on the right, plotted values are obtained using 50 control points ($NCP = 50$) (both with similar finite element meshes).

Let us recall that experimental results correspond to the Lippman equation up to a critical electrical potential u_{cr} (for the present case, the observed critical value $u_{cr} \approx 700$ V). For $u_0 > u_{cr}$, experimental results show a saturation of the wetting angle (locking phenomenon); see, e.g., [21]. As mentioned previously, the explanation of this locking phenomenon is still poorly understood by physicists. For $u_0 \approx 1050$ V, the Lippman equation predicts a total spreading of the drop on the substrate (the wetting angle vanishes).

With the present numerical model and with $NCP = 15$, we obtain a good agreement with the Lippman equation for $u_0 < 500$ V. For higher u_0 values, we do not model the angle saturation, but we observe that the contact angle is higher than the predicted value for the plane capacitor approximation.

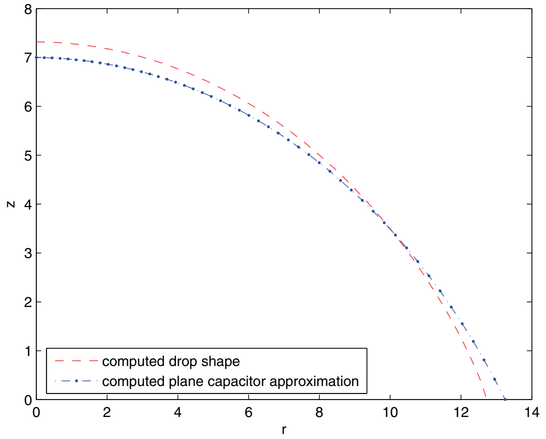


FIG. 8.6. *Computed shape compared to plane capacitor approximation shape for* $u_0 = 800$ *V.*

When increasing the number of control points to $NCP = 50$, we still obtain a good agreement with the Lippman equation for $u_0 < 400$ V. As with 15 points, we notice that the computed values are higher than the predicted values for the plane capacitor approximation. Moreover, the angle values computed with 50 points are higher than those obtained for 15 points for $u_0 > 500$ V.

Also, we compare the drop shape obtained to those obtained using the software but "forcing" the Lippman approximation (i.e., by changing σ_{LS} for each u_0 value). In Figure 8.6 the result for a drop at 400 V is presented. We again find that the wetting angle of the computed shape is higher than the Lippman predicted value.

Let us clarify that we did not manage to increase the NCP because of the well-known instability of the shape optimal design algorithms. As a matter of fact, shape optimal design algorithms become unstable when the control point density becomes similar to the finite element point density.

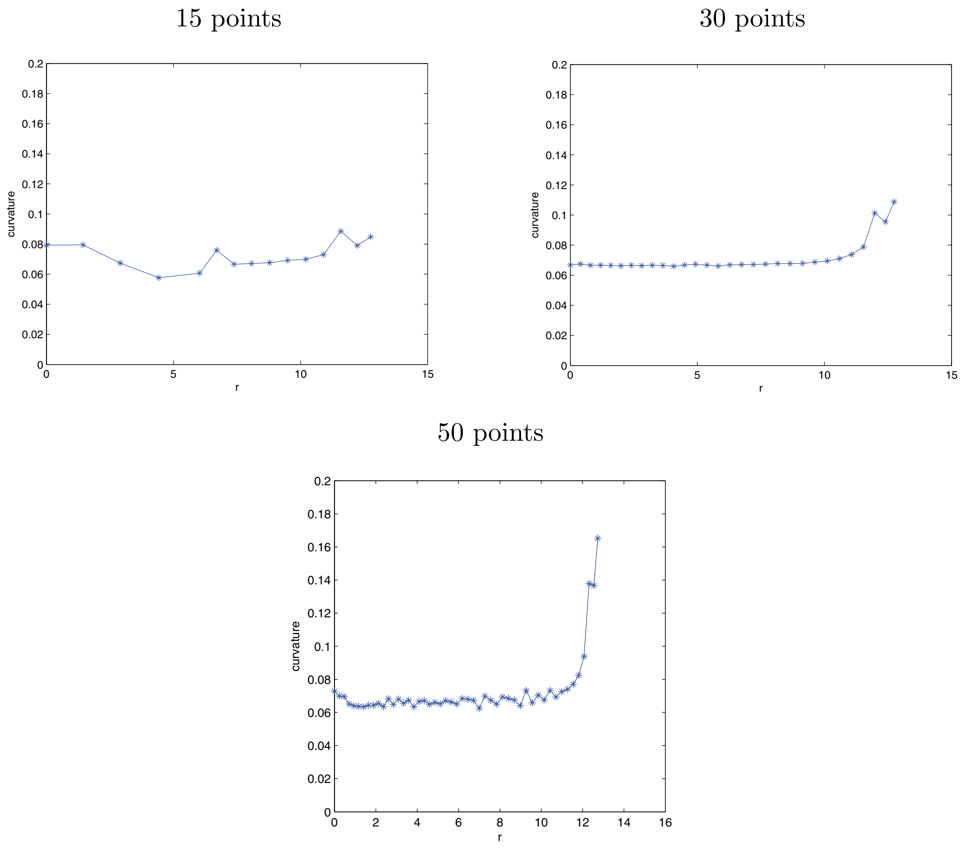
In summary, with the present model, we do not manage to properly simulate the locking phenomenon, but we do observe an overestimate of the Lippman predictions; this overestimate becomes more important when using a higher control point density.

Curvature. We use the algorithm described in the previous section; see also [14]. For all the computations we performed, the droplet shapes obtained had a constant curvature everywhere but in the vicinity of the triple point. In Figure 8.7, we present as an example (here $u_0 = 800$ V) the computed curvature at each control point. The results are presented with 15, 30, and 50 points, respectively.

In Figure 8.8, we present the curvature values for different electrical potential u_0 values with 50 control points (with curvature values corresponding to those in Figure 8.7, but for different u_0 values). In Figure 8.9, we present the gradient of the solution, i.e., the electric field.

For all computations we performed, the curvatures behave as those shown in Figure 8.8. Thus, we can make the following three main remarks:

- For the curvature, the results are more accurate with 50 points than with 15 or 30 points. With 15 or 30 points, the behavior of the curvature near the triple point appears to be less clear than with 50 points. This is due to the too small number of points near the triple line.



15 points

30 points

50 points

Fig. 8.7. *Curvature values at $u_0 = 800\,V$ for 15, 30, and 50 points, respectively.*

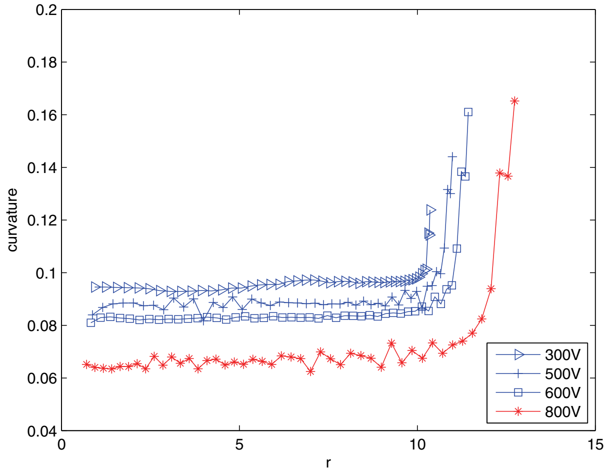


Fig. 8.8. *Curvature of the drop for several u_0 values with 50 control points.*

- For a given potential u_0, the curvature remains constant until we approach the triple line, where the curvature increases. We can see that the curvature is higher near the triple point than it is further away from it. (See Figure 8.7 for the case at 800 V. For other voltage the curvature has the same behavior; see Figure 8.8.)
- If we look at the evolution of the curvature for an increasing potential u_0, we notice the following:
  - The value of the curvature far from the triple line is constant and decreases when u_0 increases.
  - The curvature near the triple point increases, when u_0 increases. The fact that, with an increasing u_0, the curvature far from the triple point decreases is in accordance with the fact that, globally, the drop should be a portion of a sphere with an increasing radius as u_0 increases. We note that the curvature increases near the triple line; this is in accordance with the fact that the contact angle is higher than the Lippman predicted value.

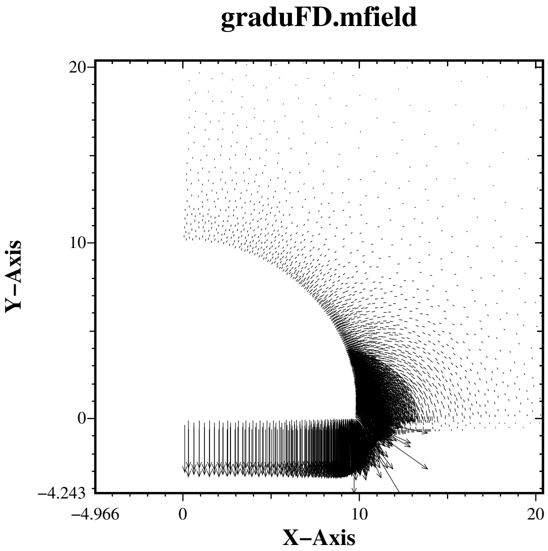


FIG. 8.9. *External electric field $\vec{E} = \vec{\nabla} u$ at $u_0 = 400$ V (zoom around the droplet).*

9. Conclusion. We have detailed and implemented a general approach for modeling the electrowetting process. The drop shape is computed as a minimum of the total energy. Our model is based on the shape optimal design methods. We test our model and software by including in the model the plane capacitor approximation (i.e., using the software with $u_0 = 0$ V and changing the value of σ_{LS} for each value of the potential). We obtain in this case an excellent agreement with the plane capacitor approximation, which contributes to validating the approach. Then, we compare numerical results obtained classically, that is to say, by changing the value of u_0, with the theoretical values for the plane capacitor approximation. In this case, the computed shapes and angles are not in agreement with this theory for a voltage higher than 300 V.

Although we did not properly obtain the locking phenomenon, the drop shape obtained deviates from the predicted shape as in [16]. Also, we did not manage to observe that the contact angle remains constant; instead, the computed contact angle values are higher than those predicted by Lippman's equation. Moreover, this overestimate becomes more important when using a higher control point density.

In other respects, we compute the curvature of the droplets. These values are globally constant except in the vicinity of the contact point where the computed curvature increases sharply. These results are in accordance with experimental results obtained in [3] and [5], which noted this increase of the curvature near the triple line.

Finally, in order to properly obtain the locking phenomenon and Young's angle at the triple line as in [5], [16], [3], further investigations based on extra singular basis functions to the finite element spaces are in progress.

Acknowledgments. The authors would like to thank C. Quilliet and M. Bienia from the Laboratory of Spectrometry Physics (LSP) in Grenoble, France, for the numerous and fruitful discussions; also, the referees for their constructive comments.

REFERENCES

[1] B. Berge, *Electrocapillarité et mouillage de films isolants par l'eau*, C. R. Acad. Sci. Sér. 2, 317 (1993), pp. 157–163.

[2] M. Bienia, C. Quilliet, and M. Vallade, *Modification of drop shape controlled by electrowetting*, Langmuir, 19 (2003), pp. 9328–9333.

[3] M. Bienia, M. Vallade, C. Quilliet, and F. Mugele, *Electrical-field-induced curvature increase on a drop of conducting liquid*, Europhys. Lett., 74 (2006), pp. 103–109.

[4] B. Bouchereau, *Modélisation et simulation numérique de l'electro-mouillage*, Ph.D. thesis, University of Grenoble I, Grenoble, France, 1997.

[5] J. Buehrle, S. Herminghaus, and F. Mugele, *Interface profiles near three-phase contact lines in electric fields*, Phys. Rev. Lett., 91 (2003), 086101.

[6] J. Céa, *Conception optimale ou identification de formes: Calcul rapide de la dérivée directionnelle de la fonction coût*, M2AN Math. Model. Numer. Anal., 20 (1986), pp. 371–402.

[7] D. Chenais, J. Monnier, and J. P. Vila, *A shape optimal design problem with convective and radiative heat transfer. Analysis and implementation*, J. Optim. Theory Appl., 110 (2001), pp. 75–117.

[8] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary Value Problems*, North–Holland, Amsterdam, 1983.

[9] B. Hamann, *Curvature approximation for triangulated surfaces*, in Geometric Modelling, G. Farin, H. Hagin, and H. N. Hemeier, eds., Springer-Verlag, Vienna, 1993, pp. 139–153.

[10] M. Meyer, M. Desbrun, P. Schroder, and A. H. Barr, *Discrete differential-geometry operators for triangulated 2-manifolds*, in Visualization and Mathematics III, H.-C. Hege and K. Polthier, eds., Springer-Verlag, Berlin, 2003, pp. 35–57.

[11] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Oxford University Press, Oxford, 2001.

[12] J. Monnier, *Shape sensitivities in a Navier-Stokes flow with convective and grey bodies radiative thermal transfer*, Optimal Control Appl. Methods, 24 (2003), pp. 237–256.

[13] J. Monnier and P. Chow-Wing-Bom, *ElectroCap: A Shape Inverse Model for an Electro-Capillary Process*, Report INRIA RR-5617, INRIA, Grenoble, France, 2005.

[14] J. Monnier, A. Benselama, and I. Cotoi, *Flow patterns in the vicinity of triple line dynamics arising from a local surface tension model*, Int. J. Multiscale Comput. Eng., 5 (2007), pp. 417–434.

[15] F. Murat and J. Simon, *Sur le Contrôle par un Domaine Géométrique*, Laboratory of Numerical Analysis, University Paris VI, Paris, 1976.

[16] A. G. Papathanasiou and A. G. Boudouvis, *Manifestation of the connection between dielectric breakdown strength and contact angle saturation in electrowetting*, Appl. Phys. Lett., 86 (2005), 164102.

[17] V. Peykov, A. Quinn, and J. Ralston, *Electrowetting: A model for contact-angle saturation*, Colloid Polym. Sci., 278 (2000), pp. 789–793.

[18] C. Quilliet and B. Berge, *Electrowetting: A recent outbreak*, Curr. Opin. Colloid Interface Sci., 6 (2001), pp. 34–39.

[19] P. Saramito, N. Roquet, and J. Etienne, *Rheolef Home Page*, http://www-lmc.imag.fr/lmc-edp/Pierre.Saramito/rheolef/, LMC-IMAG, 2002.

[20] B. Shapiro, H. Moon, R. L. Garrell, and C.-J. Kim, *Equilibrium behavior of sessile drops under surface tension, applied external fields and material variations*, J. Appl. Phys., 93 (2003), pp. 5794–5811.

[21] M. Vallet, M. Vallade, and B. Berge, *Limiting phenomena for the spreading of water on polymer films by electrowetting*, Eur. Phys. J. B, 11 (1999), pp. 583–591.

[22] H. J. J. Verheijen and M. W. J. Prins, *Reversible electrowetting and trapping of charge: Model and experiments*, Langmuir, 15 (1999), pp. 6616–6620.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1501–1519

THE RIEMANN PROBLEM FOR A NONISENTROPIC FLUID IN A NOZZLE WITH DISCONTINUOUS CROSS-SECTIONAL AREA[*]

MAI DUC THANH[†]

Abstract. We present a full investigation of the Riemann problem for a nonisentropic polytropic fluid in a nozzle with piecewise constant cross-section. First, we introduce the concept of elementary waves which turn out to make up Riemann solutions. Second, we study a procedure to select admissible stationary waves relying on the monotone criterion. By projecting all the wave curves in the (p, u)-plane, we construct Riemann solutions. Existence of Riemann solutions can be obtained for large initial data. Furthermore, we establish the uniqueness of Riemann solutions in strictly hyperbolic domains. Our argument can lead to estimate regions where the Riemann problem admits a unique solution.

Key words. gas dynamics equations, Riemann problem, conservation law, shock wave, source term, nozzle

AMS subject classifications. 35L65, 76N10, 76L05

DOI. 10.1137/080724095

1. Introduction.

In this paper we provide a full investigation of the Riemann problem for the evolution of a fluid inside a nozzle with piecewise constant cross-section. The governing equations are given by

$$\begin{aligned}&\partial_t(a\rho) + \partial_x(a\rho u) = 0,\\ &\partial_t(a\rho u) + \partial_x(a(\rho u^2 + p)) = p(\rho)\partial_x a,\\ &\partial_t(a\rho e) + \partial_x(au(\rho e + p)) = 0, \quad x \in \mathbb{R},\, t > 0.\end{aligned} \tag{1.1}$$

Here, ρ, ε, T, S, and p denote the thermodynamical variables: density, internal energy, absolute temperature, entropy, and the pressure, respectively; u is the velocity, and $e = \varepsilon + u^2/2$ is the total energy. The function $a = a(x) > 0, x \in \mathbb{R}$, is the cross-sectional area. The expression of the source term on the right-hand side of (1.1) could be understood in the sense of *nonconservative product*; see [9, 26]. To begin with, we supplement the system (1.1) with the trivial equation (see [25, 29])

$$\partial_t a = 0, \quad x \in \mathbb{R},\, t > 0. \tag{1.2}$$

That step would remove the obstacle of the source term in producing a linearly degenerate characteristic field. However, the resulting system is not strictly hyperbolic as characteristic fields coincide on certain surfaces. Therefore we will deal with the question of constructing Riemann solutions for a nonstrictly hyperbolic system. For simplicity, we assume that the fluid is polytropic ideal so that the equation of state is given by

$$p = (\gamma - 1)\rho\varepsilon, \quad 1 < \gamma \le 5/3. \tag{1.3}$$

Nevertheless, our argument can be applied for a more general class of fluids.

[*] Received by the editors May 13, 2008; accepted for publication (in revised form) November 14, 2008; published electronically March 4, 2009.
http://www.siam.org/journals/siap/69-6/72409.html

[†] Department of Mathematics, International University, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam (mdthanh@hcmiu.edu.vn).

This paper provides the constructions of all Riemann solutions of the system (1.1) for nonisentropic fluids in a nozzle with piecewise constant cross-sections for large initial data under the admissibility criterion on stationary waves. In particular, we establish the existence for large data and uniqueness in strictly hyperbolic domains. Moreover, our argument can lead to the estimates of regions where the Riemann problem admits a unique solution. This result is compatible with the uniqueness for the Cauchy problem in [16].

Recently, LeFloch and Thanh [27, 28] constructed Riemann solutions for the model of isentropic flows in a nozzle with variable cross-section and shallow water equations for arbitrary data. The Riemann problem for (1.1) was also considered by a different approach by Andrianov and Warnecke [2], where the authors introduced a new concept of solutions. For earlier work on resonant systems, see also [29, 18, 17, 11]. A careful investigation into a two-fluid model was obtained in [21]. These are typical examples of systems of balanced laws with source terms. Practically, these source terms often cause lots of inconveniences in their numerical discretization. The discretization of source terms therefore plays an important role in numerical approximations. The subject has been attracting attention of many authors for various classes of systems of balanced laws with source terms from a single conservation law, shallow water equations, or in the model of fluid flows in a nozzle with variable cross-section, to multiphase flow models; see [14, 23, 22, 15, 7, 13, 5, 6, 3, 20, 19, 32, 4, 8, 1, 10, 24, 31, 30] and the references therein. Since the system (1.1) may serve as a simple example of multiphase flow models, explicit constructions of Riemann solutions for (1.1) are interesting not only for the study of the Riemann problem itself, but also for the possibility of using these explicit solutions as references for testing various numerical schemes for multiphase flows.

The organization of this paper is as follows. In section 2 we provide general discussions and an argument for the establishment of stationary waves as elementary waves, and then we define elementary waves and determine all wave curves. Section 3 is devoted to the selection of admissible stationary waves. In section 4 we will construct Riemann solutions and establish the existence as well as the uniqueness of Riemann solutions.

2. Basic properties and elementary waves. In this section we recall basic properties of system (1.1)–(1.2) and draw elementary conclusions for stationary waves which will be used in the next sections. In particular, we derive formulas for all wave curves consisting of elementary waves.

2.1. Nonstrict hyperbolicity. To deal with the nonconservativeness of the system (1.1), we supplement it with the trivial equation (1.2). We therefore have the following system of balanced laws:

$$\begin{aligned}
&\partial_t(a\rho) + \partial_x(a\rho u) \;=\; 0,\\
&\partial_t(a\rho u) + \partial_x(a(\rho u^2 + p)) \;=\; p\partial_x a,\\
&\partial_t(a\rho e) + \partial_x(au(\rho e + p)) = 0,\\
&\partial_t a = 0, \quad x \in \mathbb{R},\, t > 0.
\end{aligned} \tag{2.1}$$

Let us take (p, S) as two independent thermodynamic variables. Then, the polytropic ideal gas equation of state can be represented by

$$\rho = \rho(p, S) = \left(\frac{p}{\gamma - 1}\mathrm{epx}\left(\frac{S_* - S}{C_v}\right)\right)^{1/\gamma}, \tag{2.2}$$

where $C_v = R/(\gamma - 1)$ and R is the specific gas constant.

A smooth solution $U = U(x,t) = (p(x,t), u(x,t), S(x,t), a(x))$ satisfies the following system of conservation laws in nonconservative form:

$$\partial_t U + A(U)\partial_x U = 0,$$

where

$$A(U) = \begin{pmatrix} u & \dfrac{\rho}{\rho_p} & 0 & \dfrac{u\rho}{a\rho_p} \\ \dfrac{1}{\rho} & u & 0 & 0 \\ 0 & 0 & u & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.3}$$

The characteristic equation is given by

$$\lambda(u-\lambda)\Big((u-\lambda)^2 - \frac{1}{\rho_p}\Big) = 0.$$

Thus, we obtain four real eigenvalues

$$\lambda_1(U) = u - c, \quad \lambda_2(U) = u, \quad \lambda_3(U) = u + c, \quad \lambda_4(U) = 0, \tag{2.4}$$

where

$$c = \frac{1}{\sqrt{\rho_p(p,S)}}, \quad \rho_p(p,S) = \frac{p^{\frac{1-\gamma}{\gamma}}}{\gamma(\gamma-1)^{\frac{1}{\gamma}}}\exp\Big(\frac{S_* - S}{\gamma C_v}\Big) > 0.$$

Obviously, we have

$$\lambda_1(U) < \lambda_2(U) < \lambda_3(U) \quad \forall U.$$

However, any of $\lambda_1(U)$, $\lambda_2(U)$, $\lambda_3(U)$ can coincide with $\lambda_4(U)$. Consequently, the system is hyperbolic but *not strictly hyperbolic*. The corresponding eigenvectors can be chosen as

$$\begin{aligned} r_1 &= \big(\rho, -\sqrt{\rho_p(p,S)}, 0, 0\big)^T, \quad r_2 = (0,0,1,0)^T, \\ r_3 &= \big(\rho, \sqrt{\rho_p(p,S)}, 0, 0\big)^T, \qquad r_4 = \Big(-\frac{2\rho(p,S)}{\rho_p(p,S)}, \frac{2}{\rho_p(p,S)}, 0, 1\Big)^T. \end{aligned}$$

Since all the eigenvalues and eigenvectors are independent of the fourth component a, investigating properties of these quantities can be reduced to the three-dimensional subspace of the coordinates (p,u,S), which is referred to as the *phase domain* of the coordinates (p,u,S). In the (p,u,S)-space, there are three surfaces, denoted by Σ_1, Σ_2, and Σ_3, on which the system fails to be strictly hyperbolic:

$$\begin{aligned} \Sigma_1 &= \{U = (p,u,S) : \lambda_4(U) = \lambda_1(U)\} = \Big\{(p,u,S) : u = \frac{1}{\sqrt{\rho_p(p,S)}}\Big\} \\ &= \Big\{(p,u,S) : u = \gamma^{\frac{1}{2}}(\gamma-1)^{\frac{1}{2\gamma}}p^{\frac{\gamma-1}{2\gamma}}\exp\Big(\frac{S_*-S}{2\gamma C_v}\Big)\Big\}, \\ \Sigma_2 &= \{U = (p,u,S) : \lambda_4(U) = \lambda_2(U)\} = \{(p,0,S)\}, \\ \Sigma_3 &= \{U = (p,u,S) : \lambda_4(U) = \lambda_3(U)\} = \Big\{(p,u,S) : u = -\frac{1}{\sqrt{\rho_p(p,S)}}\Big\} \\ &= \Big\{(p,u,S) : u = -\gamma^{\frac{1}{2}}(\gamma-1)^{\frac{1}{2\gamma}}p^{\frac{\gamma-1}{2\gamma}}\exp\Big(\frac{S_*-S}{2\gamma C_v}\Big)\Big\}. \end{aligned} \tag{2.5}$$

These surfaces, referred to as *strictly hyperbolic boundaries*, separate the phase domain into four subdomains, denoted by $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$, and $\mathcal{G}_4$, in which the system is strictly hyperbolic:

$$\begin{aligned}
&\mathcal{G}_1 = \{U = (p,u,S) : \lambda_4(U) < \lambda_1(U)\},\\
&\mathcal{G}_2 = \{U = (p,u,S) : \lambda_1(U) < \lambda_4(U) < \lambda_2(U)\},\\
&\mathcal{G}_3 = \{U = (p,u,S) : \lambda_2(U) < \lambda_4(U) < \lambda_3(U)\},\\
&\mathcal{G}_4 = \{U = (p,u,S) : \lambda_4(U) > \lambda_3(U)\}.
\end{aligned} \tag{2.6}$$

Clearly, the 2- and the 4-characteristic fields are linearly degenerate. On the other hand,

$$-\nabla\lambda_1(U)\cdot r_1(U) = \nabla\lambda_3(U)\cdot r_3(U) = \frac{\gamma+1}{2}\sqrt{\rho_p} > 0 \quad \forall U, \tag{2.7}$$

which implies that the 1- and the 3-characteristic fields are genuinely nonlinear.

2.2. Stationary smooth solutions. *A stationary smooth solution* U of (1.1) is a time-independent smooth solution. Thus, the derivative with respect to t in (1.1) can be omitted. Stationary solutions of the initial value problem for (1.1) are, therefore, the ones for the following ordinary differential equations:

$$\begin{aligned}
&(a\rho u)' = 0,\\
&(a(\rho u^2 + p))' = pa',\\
&(au(\rho e + p))' = 0,
\end{aligned} \tag{2.8}$$

where $(.)'$ stands for $d(.)/dx$.

We will show that the specific entropy is conserved across stationary waves. Therefore, for those concerning stationary waves, the choices of thermodynamic independent variables (ρ, S) or (p, S) are equivalent, since $p = p(\rho, S_0)$ implies $\rho = \rho(p, S_0)$, and vice versa. To make calculations simpler, let us, however, choose the thermodynamic independent variables in this subsection to be (ρ, S). Solutions of (2.8) are subject to the initial condition

$$(\rho, u, S, a)(x_0) = (\rho_0, u_0, S_0, a_0). \tag{2.9}$$

The *specific enthalpy* is given by

$$h \,=\, \varepsilon + pv, \tag{2.10}$$

which satisfies

$$dh \,=\, T\,dS + v\,dp.$$

The following lemma enables us to calculate stationary waves.

Lemma 2.1. *For smooth solutions, the system* (2.8) *is equivalent to*

$$\begin{aligned}
&(a\rho u)' = 0,\\
&\Big(\frac{u^2}{2} + h(\rho, S)\Big)' = 0,\\
&S' = 0.
\end{aligned} \tag{2.11}$$

Proof. The first equation of (2.11) can be expressed as

$$a\rho u = a_0\rho_0 u_0 = C,$$

where C is a constant. Thus, the second equation can be written as

$$(C \cdot u + a \cdot p)' = p \cdot a'$$

or

$$C \cdot u' + a \cdot p' = 0.$$

This yields

$$uu' + \frac{p'}{\rho} = uu' + p'v = 0, \quad v = \frac{1}{\rho}. \tag{2.12}$$

Now, provided $C \neq 0$, the third equation of (2.8) can be written as

$$\varepsilon' + uu' + (pv)' = 0. \tag{2.13}$$

Recall the thermodynamic identity that

$$TdS = d\varepsilon + pdv.$$

Since we are considering stationary waves, i.e., solutions independent of time, the thermodynamic identity applied to this kind of wave gives

$$\varepsilon' = TS' - pv'.$$

Substituting this into (2.13), we get

$$TS' + p'v + uu' = 0,$$

or, from (2.12), it holds that

$$S' = 0.$$

Since $S' = 0$, we have

$$p'v = h'(\rho, S),$$

and from (2.12), this yields the second line of (2.11). Lemma 2.1 is completely proved. ▯

2.3. Stationary contact waves. Suppose now we have a discontinuity with propagation speed λ. Then, the Rankine–Hugoniot relation associated with (1.2) gives

$$-\lambda[a] = 0, \tag{2.14}$$

where $[a] := a_+ - a_-$ is the jump of the cross-section a. Thus, (2.14) implies that there are two possibilities:

(i) either $\lambda = 0$: the shock speed vanishes,

(ii) or $[a] = 0$: the component a remains constant across the shock.

First, suppose (i) that the component a is discontinuous and, therefore, the shock speed vanishes. The solution is independent of time, and it is natural to search for a solution as the limit of a sequence of time-independent smooth solutions, or stationary solutions of (1.1)–(1.2), which are given by (2.11). The integral curve of (2.11) passing through each point (ρ_0, u_0, S_0, a_0) can be parameterized by ρ, say

$$\rho \mapsto (\rho, u(\rho), S_0, a(\rho)),$$

and satisfies

$$\begin{aligned}
&a\rho u = a_0\rho_0 u_0,\\
&\frac{u^2}{2} + h(\rho, S) = \frac{u_0^2}{2} + h(\rho_0, S_0),\\
&S = S_0.
\end{aligned}$$

Letting $\rho \to \rho_1$ and setting $u_1 = u(\rho_1), a_1 = a(\rho_1)$, we see that the states (ρ_0, u_0, S_0, a_0), (ρ_1, u_1, S_0, a_1) satisfy the Rankine–Hugoniot relations

$$\begin{aligned}
&[a\rho u] = 0,\\
&\left[\frac{u^2}{2} + h(\rho, S_0)\right] = 0,\\
&[S] = 0.
\end{aligned}$$

So

$$\begin{aligned}
&u = \frac{a_0\rho_0 u_0}{a\rho},\\
&\frac{u^2}{2} + h(\rho, S_0) = \frac{u_0^2}{2} + h(\rho_0, S_0),\\
&S = S_0.
\end{aligned} \tag{2.15}$$

Set $h_0 := h(\rho_0, S_0)$; then the relations (2.15) define a curve $\mathcal{W}_4(U_0)$ in the phase domain. Precisely, $\mathcal{W}_4(U_0)$ is the intersection of the two surfaces defined by the first and the second equations in (2.15) in the (ρ, u, a)-space (entropy is constant). Since the mapping $\rho \mapsto p(\rho, S_0)$ is monotone increasing, in the (p, u)-plane this curve can be parameterized as $p \mapsto u(p)$, which is monotone decreasing for $u_0 > 0$ and monotone increasing for $u_0 < 0$.

The system (2.15) defines a function $a \mapsto (\rho = \rho(a; U_0), p = p(a; U_0), u = u(a; U_0))$, where $\rho = \rho(a; U_0)$ is determined by

$$\Phi(\rho, a; U_0) := -\frac{2\kappa\gamma}{\gamma - 1}\rho^{\gamma+1} + \Big(u_0^2 + \frac{2\kappa\gamma}{\gamma - 1}\rho_0^{\gamma-1}\Big)\rho^2 - \Big(\frac{a_0 u_0 \rho_0}{a}\Big)^2 = 0, \tag{2.16}$$

where

$$\kappa = A(S_0), \quad A(S) = (\gamma - 1)\exp\Big(\frac{S - S_*}{C_v}\Big), \tag{2.17}$$

and then u is given by the first equation of (2.15), $S = S_0$, and $p = p(\rho, S_0)$.

2.4. Shocks, rarefaction waves, and contact discontinuities. Let us now suppose (ii) that the component a remains constant on both sides of the discontinuity. Eliminating a from (1.1), we obtain the usual gas dynamics equations

$$\begin{aligned}&\partial_t\rho+\partial_x(\rho u)=0,\\&\partial_t(\rho u)+\partial_x(\rho u^2+p)=0,\\&\partial_t(\rho e)+\partial_x(u(\rho e+p))=0.\end{aligned}\tag{2.18}$$

Thus, all elementary waves of the system (2.18) are obtained in a usual way which can be found in any standard material on gas dynamics equations. In particular, the reader is referred to [12]. However, it is convenient to brief basic facts of these waves here.

The Rankine–Hugoniot relations corresponding to (2.18) are given by

$$\begin{aligned}&-\lambda[\rho]+[\rho u]=0,\\&-\lambda[\rho u]+[\rho u^2+p(\rho)]=0,\\&-\lambda[\rho e]+[u(\rho e+p)]=0,\end{aligned}\tag{2.19}$$

where $[\rho]=\rho_1-\rho_0$, $[\rho u]=\rho_1u_1-\rho_0u_0$, etc., and λ is the speed of propagation of the discontinuity connecting the states U_0 and U_1.

Then the Hugoniot set issuing from a given state U_0 consisting of all states U that can be connected to U_0 by a discontinuity satisfying the Rankine–Hugoniot relations is determined by

$$\begin{aligned}&M:=\rho_0(u_0-\lambda)=\rho_1(u_1-\lambda),\\&\rho_0(u_0-\lambda)^2+p_0=\rho_1(u_1-\lambda)^2+p_1,\\&\left(\rho_0\left(\varepsilon_0+\frac{(u_0-\lambda)^2}{2}\right)+p_0\right)=\left(\rho_1\left(\varepsilon_1+\frac{(u_1-\lambda)^2}{2}\right)+p_1\right).\end{aligned}\tag{2.20}$$

If $M=0$, then

$$\begin{aligned}&u_0=\lambda=u_1,\\&p_0=p_1.\end{aligned}\tag{2.21}$$

Thus, the discontinuity is exactly the 2-contact discontinuity corresponding to $\lambda=\lambda_2=u$. When $M\neq 0$, we obtain a 1-shock if $M>0$, and a 3-shock if $M<0$.

Next, all discontinuities of (2.19) associated with nonlinear characteristic fields are required to satisfy the *Lax shock inequalities*

$$\lambda_i(U_0)>\lambda(U_0,U)>\lambda_i(U),\quad i=1,3,\tag{2.22}$$

where $\lambda(U_0,U)$ is the shock speed of the shock connecting the left-hand state U_0 to the right-hand state U belonging to the Hugoniot set issuing from U_0.

For polytropic ideal gas (1.3), the Lax shock inequalities (2.22) yield the following:

(a) For a 1-shock, the Lax shock inequalities are equivalent to

$$\rho_1\geq\rho_0,\quad p_1\geq p_0,\quad S_1\geq S_0,\quad u_1\leq u_0.\tag{2.23}$$

(b) For a 3-shock, the Lax shock inequalities are equivalent to

$$\rho_1\leq\rho_0,\quad p_1\leq p_0,\quad S_1\leq S_0,\quad u_1\geq u_0.\tag{2.24}$$

From (2.20) and (2.23), we obtain the first *forward* shock curve $\mathcal{S}_1(U_0)$ issuing from U_0 consisting of all right-hand states that can be connected to a given left-hand state U_0 as

$$\begin{aligned}\mathcal{S}_1(U_0): \qquad & v = v_1(U_0,p) = \frac{v_0(\mu p + p_0)}{p+\mu p_0}, \qquad \text{where} \quad \mu = \frac{\gamma-1}{\gamma+1},\\ & u = u_1(U_0,p) = u_0 - (p-p_0)\sqrt{\frac{(1-\mu)v_0}{p+\mu p_0}}, \quad p \geq p_0.\end{aligned} \tag{2.25}$$

From (2.20) and (2.24), we also obtain the third *backward* shock curve $\mathcal{S}_3(U_0)$ issuing from U_0 consisting of all left-hand states that can be connected to a given right-hand state U_0. Then, the inequalities in (2.24) must be reversed. So $\mathcal{S}_3(U_0)$ is given by

$$\begin{aligned}\mathcal{S}_3(U_0): \qquad & v = v_3(U_0,p) = \frac{v_0(\mu p + p_0)}{p+\mu p_0},\\ & u = u_3(U_0,p) = u_0 + (p-p_0)\sqrt{\frac{(1-\mu)v_0}{p+\mu p_0}}, \quad p \geq p_0.\end{aligned} \tag{2.26}$$

Next, rarefaction waves in genuinely nonlinear characteristic fields corresponding to λ_1 and λ_3 are continuous piecewise smooth self-similar solutions of (1.1) of the form

$$U(x,t) = V(\xi), \quad \xi = x/t.$$

Recall that the *forward* 1-rarefaction curve $\mathcal{R}_1(U_0)$ consisting of all right-hand states U that can be connected to the left-hand state U_0 by a 1-rarefaction wave is given by

$$\begin{aligned}\mathcal{R}_1(U_0): \qquad u = u_1(U_0,p) &= u_0 - \int_{p_0}^{p} \frac{\sqrt{\rho_p(z,S_0)}}{\rho(z,S_0)}dz,\\ &= u_0 - \frac{2\gamma^{1/2}}{(\gamma-1)^{1-1/2\gamma}}\exp\Big(\frac{S_0-S_*}{2C_v\gamma}\Big)(p^{(\gamma-1)/2\gamma} - p_0^{(\gamma-1)/2\gamma}).\end{aligned} \tag{2.27}$$

Similarly, the *backward* 3-rarefaction curve $\mathcal{R}_3(U_0)$ consisting of all left-hand states U that can be connected to the right-hand state U_0 by a 1-rarefaction wave is given by

$$\begin{aligned}\mathcal{R}_3(U_0): \qquad u = u_3(U_0,p) &= u_0 + \int_{p_0}^{p} \frac{\sqrt{\rho_p(z,S_0)}}{\rho(z,S_0)}dz,\\ &= u_0 + \frac{2\gamma^{1/2}}{(\gamma-1)^{1-1/2\gamma}}\exp\Big(\frac{S_0-S_*}{2C_v\gamma}\Big)(p^{(\gamma-1)/2\gamma} - p_0^{(\gamma-1)/2\gamma}).\end{aligned} \tag{2.28}$$

The wave curves associated with the genuinely nonlinear characteristic fields are then defined as

$$\mathcal{W}_i(U_0) := \mathcal{S}_i(U_0) \cup \mathcal{R}_i(U_0), \quad i = 1,3.$$

It is not difficult to check that the wave curve $\mathcal{W}_1(U_0)$ projected in (p,u)-plane, as a function $p \mapsto u$, is continuous, monotone decreasing, and the wave curve $\mathcal{W}_3(U_0)$ projected in (p,u)-plane, as a function $p \mapsto u$, is continuous, monotone increasing. In conclusion, we obtain the same waves associated with the first, second, and the third characteristic fields as in the usual gas dynamics equations.

The above arguments allow us to define elementary waves of the system (1.1) which form Riemann solutions.

Definition 2.2. *Elementary waves for the system* (3.1) *are the following ones:*

- *Lax shocks, rarefaction waves, and contact discontinuities of the usual gas dynamics equations corresponding to the case a is constant in* (1.1)*.*
- Stationary contacts *with zero propagation speed are given by* (2.15)*.*

3. Selection of admissible stationary waves. As seen in the previous section, the density ρ across a stationary wave with a given state U_0 is determined as zeros depending on the parameter a of the function

$$\Phi(\rho, a; U_0) := -\frac{2\kappa\gamma}{\gamma-1}\rho^{\gamma+1} + \Big(u_0^2 + \frac{2\kappa\gamma}{\gamma-1}\rho_0^{\gamma-1}\Big)\rho^2 - \Big(\frac{a_0u_0\rho_0}{a}\Big)^2, \tag{3.1}$$

where

$$\kappa = A(S_0), \quad A(S) = (\gamma-1)\text{exp}\Big(\frac{S-S_*}{C_v}\Big).$$

If $u_0 = 0$, then the equation $\Phi(\rho, a; U_0) = 0$ gives three roots; therefore there are three states $(\rho_0, 0, S_0), (0, \pm\sqrt{(2\kappa\gamma)/(\gamma-1)}\rho_0^{(\gamma-1)/2}, S_0)$ that can be connected to U_0 by a stationary wave. Assume $u_0 \neq 0$. First, observe that since we look for zeros of the function, we just consider those values ρ such that

$$-\frac{2\kappa\gamma}{\gamma-1}\rho^{\gamma+1} + \Big(u_0^2 + \frac{2\kappa\gamma}{\gamma-1}\rho_0^{\gamma-1}\Big)\rho^2 \geq 0,$$

which requires

$$\rho \leq \bar{\rho}(U_0) := \Big(\frac{\gamma-1}{2\kappa\gamma}u_0^2 + \rho_0^{\gamma-1}\Big)^{\frac{1}{\gamma-1}}. \tag{3.2}$$

Thus, we need to investigate the function Φ on the interval $[0, \bar{\rho}(U_0)]$ only. We have

$$\frac{d\Phi(U_0, \rho; a)}{d\rho} = -2\kappa\gamma\frac{\gamma+1}{\gamma-1}\rho^{\gamma} + 2\Big(u_0^2 + \frac{2\kappa\gamma}{\gamma-1}\rho_0^{\gamma-1}\Big)\rho$$

so that

$$\begin{aligned} &\frac{d\Phi(\rho; a, U_0)}{d\rho} > 0, \quad \rho < \rho_{\max}(U_0),\\ &\frac{d\Phi(U_0, \rho; a)}{d\rho} < 0, \quad \rho > \rho_{\max}(U_0), \end{aligned} \tag{3.3}$$

where

$$\rho_{\max}(U_0) := \Big(\frac{\gamma-1}{\kappa\gamma(\gamma+1)}u_0^2 + \frac{2}{\gamma+1}\rho_0^{\gamma-1}\Big)^{\frac{1}{\gamma-1}}. \tag{3.4}$$

Moreover,

$$\Phi(\rho = 0; a, U_0) = \Phi(\rho = \bar{\rho}; a, U_0) = -\Big(\frac{a_0u_0\rho_0}{a}\Big)^2 < 0, \quad u_0 \neq 0. \tag{3.5}$$

From (3.4) and (3.5), we can see that Φ admits a zero if and only if

$$\Phi(\rho = \rho_{\max}; a, U_0) \geq 0.$$

Equivalently

$$a \geq a_{\min}(U_0) := \frac{a_0 \rho_0 |u_0|}{\sqrt{\kappa\gamma}\rho_{\max}^{\frac{\gamma+1}{2}}(\rho_0, u_0)}. \tag{3.6}$$

If $a > a_{\min}(U_0)$, then there are exactly two values $\rho_*(U_0, a) < \rho_{\max}(U_0) < \rho^*(U_0, a)$ such that

$$\Phi(\rho_*(U_0, a); a, U_0) = \Phi(\rho^*(U_0, a); a, U_0) = 0.$$

By considering (p, S) as the two independent thermodynamic variables and phase domain in the (p, u, S, a) space, we have the following lemma.

Lemma 3.1 (stationary waves). *There exists a stationary contact from a given state $U_0 = (p_0, u_0, S_0, a_0)$ connecting to some state $U = (p, u, S = S_0, a)$ if and only if $a \geq a_{\min}(U_0)$. More precisely, we have the following:*

(i) *If $a < a_{\min}(U_0)$, there are no stationary contacts.*

(ii) *If $a \geq a_{\min}(U_0)$ along the curve $\mathcal{W}_4(U_0)$, there are exactly two points $U_* := (p = p(\rho_*(U_0, a), S_0), u = a_0\rho_0 u_0/(a\rho_*(U_0, a)), S = S_0, a)$, and $U^* := (p = p(\rho^*(U_0, a), S_0), u = a_0\rho_0 u_0/(a\rho^*(U_0, a)), S = S_0, a)$, where $\rho_*(U_0, a) < \rho_{\max}(U_0) < \rho^*(U_0, a)$ satisfying*

$$\Phi(\rho_*(U_0, a); a, U_0) = \Phi(\rho^*(U_0, a); a, U_0) = 0. \tag{3.7}$$

These two states U_*, U^* coincide only if $a = a_{\min}(U_0)$.

Since the function $\rho \mapsto p(\rho, S_0)$ is monotone, the following lemma can be obtained directly from Lemma 2.3 [27], which provides some useful properties of the above quantities. Setting

$$p_{\max} = p(\rho_{\max}, S_0), \quad p_* = p(\rho_*, S_0), \quad p^* = p(\rho^*, S_0),$$

and using notation in Lemma 3.1, we have the following lemma.

Lemma 3.2. *Given a state $U_0 = (p_0, u_0, S_0, a_0)$, we have the following.*

(a) *It holds that*

$$\begin{aligned} p_{\max}(U_0) &< p_0, \quad U_0 \in \mathcal{G}_2 \cup \mathcal{G}_3, \\ p_{\max}(U_0) &> p_0, \quad U_0 \in \mathcal{G}_1 \cup \mathcal{G}_4. \end{aligned}$$

(b) *The state $U_* = (p = p(\rho_*(U_0, a), S_0), u = a_0\rho_0 u_0/(a\rho_*(U_0, a)), S = S_0, a) \in \mathcal{G}_1$ if $u_0 < 0$, and $U_* \in \mathcal{G}_4$ if $u_0 > 0$. The state $U^* := (p = p(\rho^*(U_0, a), S_0), u = a_0\rho_0 u_0/(a\rho^*(U_0, a)), S = S_0, a) \in \mathcal{G}_2$ if $u_0 < 0$, and $U^* \in \mathcal{G}_3$ if $u_0 > 0$. In addition, we have the following:*

- *If $a > a_0$, then*

$$p_*(U_0, a) < p_0 < p^*(U_0, a).$$

- *If $a < a_0$, then*

$$\begin{aligned} p_0 &< p_*(U_0, a) \quad \textit{for} \quad U_0 \in \mathcal{G}_1 \cup \mathcal{G}_4, \\ p_0 &> p^*(U_0, a) \quad \textit{for} \quad U_0 \in \mathcal{G}_2 \cup \mathcal{G}_3. \end{aligned}$$

(c)

$$\begin{aligned} &a_{\min}(U,a) < a, \quad U \in \mathcal{G}_i,\ i=1,2,3,4,\\ &a_{\min}(U,a) = a, \quad U \in \Sigma_1 \cup \Sigma_3,\\ &a_{\min}(U,a) = 0, \quad p=0 \quad \textit{or} \quad u=0. \end{aligned}$$

In combining waves, we must know when the order of wave speeds associated with different characteristic fields changes. Precisely, we want to know when the shock speeds in the nonlinear characteristic fields equal zero. From (2.20), for polytropic gas (1.3), we can calculate the shock speeds as

$$\begin{aligned} \lambda = \lambda(U_0,U) &= \frac{\rho u - \rho_0 u_0}{\rho - \rho_0}\\ &= \frac{\rho(u-u_0) + u_0(\rho-\rho_0)}{\rho-\rho_0}\\ &= u_0 - v_0\frac{u-u_0}{v-v_0}\\ &= \begin{cases} u_0 - v_0\sqrt{\frac{p+\mu p_0}{(1-\mu)v_0}} & \text{for } 1-\text{(forward) shocks},\\ u_0 + v_0\sqrt{\frac{p+\mu p_0}{(1-\mu)v_0}} & \text{for } 3-\text{(forward) shocks}. \end{cases} \end{aligned}$$

Thus, the 1-shock speed $\lambda = \bar{\lambda}_1(U_0,U)$ from a given left-hand state U_0 to a right-hand state U on the Hugoniot set issuing from U_0 vanishes if

$$\begin{aligned} &u_0 > 0,\\ &p = \tilde{p}_0 := \frac{(1-\mu)u_0^2}{v_0} - \mu p_0. \end{aligned}$$

Besides, the Lax shock inequalities require that $\tilde{p}_0 > p_0$. This means

$$\frac{(1-\mu)u_0^2}{v_0} - \mu p_0 > p_0$$

or

$$u_0^2 > \frac{1+\mu}{1-\mu}p_0 v_0 = \gamma p_0 v_0 = \frac{1}{\rho_p(p_0,S_0)} = c^2.$$

Since $u_0 > 0$, this is equivalent to

$$\bar{\lambda}_1(U_0,U) = u_0 - c > 0,$$

which says that $U_0 \in \mathcal{G}_1$. Similarly, the backward 3-shock speed $\bar{\lambda}_3(U_0,U)$ from a given right-hand state U_0 to a left-hand state U vanishes if $U_0 \in \mathcal{G}_4$ and $p = \frac{(1-\mu)u_0^2}{v_0} - \mu p_0$. We therefore arrive at the following proposition.

PROPOSITION 3.3. (a) *The* 1-*shock speed* $\bar{\lambda}_1(U_0,U)$, *(for* $p > p_0$*) may change sign along the forward* 1-*shock curve* $\mathcal{S}_1(U_0)$. *More precisely, we have the following:*

(i) *If* $U_0 \in \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4$, *then* $\bar{\lambda}_1(U_0,U)$ *remains negative:*

$$\bar{\lambda}_1(U_0,U) < 0, \quad U \in \mathcal{S}_1(U_0).$$

(ii) *If* $U_0 \in \mathcal{G}_1$*, then* $\bar\lambda_1(U_0, U)$ *vanishes once at some point* $U = \tilde U_0 \in \mathcal{G}_2$ *corresponding to a value* $p = \tilde p_0 = \frac{(1-\mu)u_0^2}{v_0} - \mu p_0$ *on the* 1*-shock curve* $\mathcal{S}_1(U_0)$ *such that*

$$\begin{aligned}&\bar\lambda_1(U_0, \tilde U_0) = 0,\\ &\bar\lambda_1(U_0, U) > 0, \quad p \in (p_0, \tilde p_0),\\ &\bar\lambda_1(U_0, U) < 0, \quad p \in (\tilde p_0, +\infty).\end{aligned} \tag{3.8}$$

(b) *The* 3*-shock speed* $\bar\lambda_3(U_0, U)$ *may change sign along the backward* 3*-shock curve* $\mathcal{S}_3(U_0)$ $(p > p_0)$*. More precisely,*

(i) *If* $U_0 \in \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$*, then* $\bar\lambda_1(U_0, U)$ *remains positive:*

$$\bar\lambda_3(U_0, U) > 0, \quad U \in \mathcal{S}_3(U_0).$$

(ii) *If* $U_0 \in \mathcal{G}_4$*, then* $\bar\lambda_3(U_0, U)$ *vanishes once at some point* $U = \tilde U_0 \in \mathcal{G}_3$ *corresponding to a value* $p = \tilde p_0 = \frac{(1-\mu)u_0^2}{v_0} - \mu p_0$ *on the backward* 3*-shock curve* $\mathcal{S}_3(U_0)$ *such that*

$$\begin{aligned}&\bar\lambda_3(U_0, \tilde U_0) = 0,\\ &\bar\lambda_3(U_0, U) < 0, \quad p \in (p_0, \tilde p_0),\\ &\bar\lambda_3(U_0, U) > 0, \quad p \in (\tilde p_0, +\infty).\end{aligned} \tag{3.9}$$

As shown in [27], the Riemann problem for (1.1)–(1.2) may admit up to a one-parameter family of solutions. This phenomenon can be avoided by requiring Riemann solutions to satisfy a monotone condition on the component a. Motivated by [27], we impose the following criterion on stationary waves of (1.1).

Admissibility Criterion 3.1. *Along the stationary curve in the* (ρ, u)*-plan between left- and right-hand states of any stationary wave, the component* a *obtained from* (2.15) *and expressed as a function of* ρ *has to be monotone in* ρ*.*

Since the specific entropy is constant along stationary curves, the following lemma can be established as in the case of isentropic gases (see [27]); therefore, we omit the proof.

Lemma 3.4. *Admissibility Criterion* 3.1 *is equivalent to the statement that any stationary wave has to remain in the closure of only one domain* $\mathcal{G}_i$*,* $i = 1, 2, 3, 4$*.*

Lemma 3.4 implies that if $U_0 \in \mathcal{G}_1 \cup \mathcal{G}_4$, then $p_*(U_0, a)$ is used, while $p^*(U_0, a)$ is used when $U_0 \in \mathcal{G}_2 \cup \mathcal{G}_3$.

4. Existence and uniqueness of the Riemann problem. In this section we will establish global existence and uniqueness of the Riemann problem for (1.1)–(1.2). Without loss of generality (by changing coordinates $x \mapsto -x, \quad u \mapsto -u$, if necessary), we can assume for definitiveness in this section that

$$a_L < a_R.$$

To construct Riemann solutions of (1.1)–(1.2), we project all the wave curves on the (p, u)-plane. Moreover, we will use the following notation:

(i) $W_k(U_i, U_j)$ $(S_k(U_i, U_j), R_k(U_i, U_j))$ denotes the kth-wave (kth-shock, kth-rarefaction wave, respectively) connecting the left-hand state U_i to the right-hand state U_j.

(ii) $W_m(U_i, U_j) \oplus W_n(U_j, U_k)$ indicates that there is an mth-wave from the left-hand state U_i to the right-hand state U_j, followed by an nth-wave from the left-hand state U_j to the right-hand state U_k.

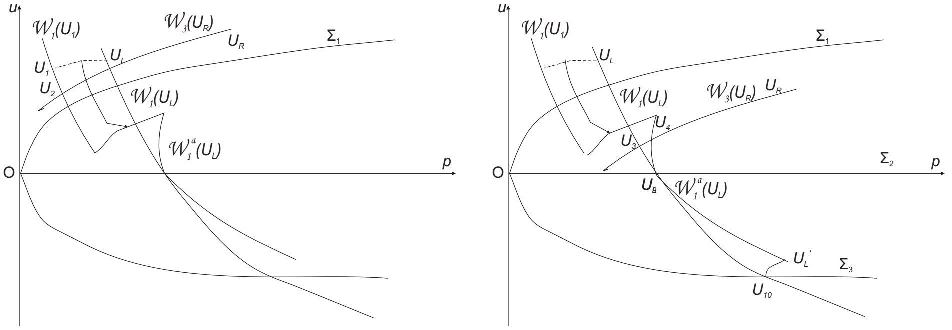


FIG. 4.1. *Riemann solution with structures* (4.1) *(left) and* (4.4) *(right).*

(iii) $\bar{U}$ denotes the right-hand state resulted by a contact discontinuity from a left-hand state U associated with the second characteristic field $\lambda_2 = u$. Observe that in the (p, u)-plane, U and $\bar{U}$ share the same location.

(iv) $\hat{U}$ denotes the right-hand state resulted by a contact discontinuity from a left-hand state U associated with the fourth characteristic field $\lambda_4 = 0$.

4.1. Explicit solutions and uniqueness for $\boldsymbol{U_L \in \mathcal{G}_1}$.

Construction N1. This construction holds for U_R, belongs to $\mathcal{G}_1 \cup \Sigma_1$, and belongs to some part of $\mathcal{G}_2$. Let U_1 be the state obtained by jumping from U_L by a stationary contact from the level a_L to the level a_R. Whenever $\mathcal{W}_1(U_1) \cap \mathcal{W}_3(U_R) \neq \emptyset$, there is a solution defined as follows. Let

$$\{U_2\} = \mathcal{W}_1(U_1) \cap \mathcal{W}_3(U_R).$$

Then the solution is

$$W_4(U_L, U_1) \oplus W_1(U_1, U_2) \oplus W_2(U_2, \bar{U}_2) \oplus W_3(\bar{U}_2, U_R). \tag{4.1}$$

The construction makes sense if $\lambda_1(U_1, U_2) \geq 0$. See Figure 4.1(left).

If $\mathcal{W}_1(U_1) \cap \mathcal{W}_3(U_R) = \emptyset$, then there is a vacuum. In fact, let $\{M\} = \mathcal{W}_1(U_1) \cap \{p = 0\}$, $\{N\} = \mathcal{W}_3(U_R) \cap \{p = 0\}$. The solution is

$$W_4(U_L, U_1) \oplus W_1(U_1, M) \oplus W_1(M, N) \oplus W_3(M, U_R). \tag{4.2}$$

Construction N2. This construction holds for U_R in $\mathcal{G}_2 \cup G_3 \cup \Sigma_1 \cup \Sigma_2$ and some part of $\mathcal{G}_1$. Consider the wave curve $\mathcal{W}_1(U_L)$. Let $\tilde{U}_L \in \mathcal{W}_1(U_L) \cap \mathcal{G}_2$ be the state at which the shock speed vanishes, i.e., $\lambda_1(U_L, \tilde{U}_L) = 0$, in view of Proposition 3.3. Let $\tilde{U}_L = (\tilde{p}_L, \tilde{u}_L)$. Then, from any point $U = (p, u) \in \mathcal{W}_1(U_L), p \geq \tilde{p}_L$, which means U is positioned "lower" than $\tilde{U}_L$, a stationary wave jumps from U from $a = a_L$ to $a = a_R$ to some state $\bar{U}$ using $W_4(U, \bar{U})$. These states $\bar{U}$ form a curve U which varies along $\mathcal{W}_1(U_L)$ "downward" from $\tilde{U}_L$. Precisely, set the "composite" curve

$$\mathcal{W}_1^a(U_L) := \{\bar{U} : \exists W_4(U, \bar{U}) \quad \text{from } a_L \quad \text{to } a_R, U = (p, u) \in \mathcal{W}_1(U_L), p \geq \tilde{p}_L\}. \tag{4.3}$$

Whenever $\mathcal{W}_3(U_R) \cap \mathcal{W}_1^a \neq \emptyset$, there will be a Riemann solution. In fact, let $\mathcal{W}_3(U_R) \cap \mathcal{W}_1^a = \{U_4\}$ and U_3 be the point on $\mathcal{W}_1(U_L)$ that corresponds to the stationary wave $W_4(U_3, U_4)$ or $W_4(\bar{U}_3, U_4)$. Then, the solution can be

$$S_1(U_L, U_3) \oplus W_4(U_3, U_4) \oplus W_2(U_4, \bar{U}_4) \oplus W_3(\bar{U}_4, U_R) \tag{4.4}$$

if $u_3 \geq 0$, and

$$S_1(U_L, U_3) \oplus W_2(U_3, \bar{U}_3) \oplus W_4(\bar{U}_3, U_4) \oplus W_3(U_4, U_R) \tag{4.5}$$

if $u_3 < 0$ and $\lambda_3(U_4, U_R) \geq 0$. See Figure 4.1(right).

Construction N3. This construction shows a connection between Constructions N1 and N2. Here, we meet an interesting phenomenon when wave speeds associated with different characteristic fields coincide. Precisely, there are solutions containing three waves with the same zero speed. This can be seen as follows. From U_L, the solution jumps by a stationary wave $W_4(U_L, A := \hat{U}_L(a))$ with an intermediate value of cross-section $a \in [a_L, a_R]$. Then, from A, the solution jumps to some state $B := \tilde{A} \in \mathcal{G}_2$ using $S_1(A, B)$ with $\lambda_1(A, B) = 0$. Next, the solution jumps from B to some state $C = U(a) := \hat{B}$ using a stationary wave $W_4(B, U(a))$ to shift the cross-section a to a_R. It is not difficult to check that the mapping

$$[a_L, a_R] \ni a \mapsto U(a) \tag{4.6}$$

is locally Lipschitz with a deterministic Lipschitz constant K on any compact neighborhood of U_L. Set

$$\mathcal{L}(U_L, a_R) = \{U(a) | a \in [a_L, a_R]\}.$$

Whenever

$$\mathcal{W}_3(U_R) \cap \mathcal{L}(U_L; a_R) \neq \emptyset$$

there is a solution containing three discontinuities having the same speed zero. Precisely, the solution begins with a stationary 4-wave from U_L to A, followed by a 1-shock with zero speed from A to B, then followed by a stationary 4-wave from B to $U(a)$, since $u(a) > 0$. The solution continues with a 2-wave from $U(a)$ to some state $D := \bar{U}(a)$, and then it arrives at U_R using a 3-wave. We therefore have a solution of the form

$$W_4(U_L, A) \oplus S_1(A, B) \oplus W_4(B, U(a)) \oplus W_2(U(a), \bar{U}(a)) \oplus W_3(\bar{U}(a), U_R). \tag{4.7}$$

See Figure 4.2(left).

Construction N4: $\boldsymbol{U_R \in \mathcal{G}_4}$. This construction can be applied for arbitrary U_L. For $U_R \in \mathcal{G}_4$, there is a stationary wave from U_R with $a = a_R$ to some state $U_8 \in \mathcal{G}_4$ with $a = a_L$. Let U_7 be the intersection point of $\mathcal{W}_3(U_8)$ and $\mathcal{W}_1(U_L)$. If $U_7 \in \mathcal{G}_4$, then there is a Riemann solution of the form

$$S_1(U_L, U_7) \oplus W_2(U_7, \bar{U}_7) \oplus W_3(\bar{U}_7, U_8) \oplus W_4(U_8, U_R). \tag{4.8}$$

See Figure 4.2 (right).

Construction N5: This construction conditionally holds for $\boldsymbol{U_R \in \mathcal{G}_4}$. This construction also holds for $\mathcal{W}_1(U_L) \cap \mathcal{G}_3 \neq \emptyset$. There is some set Δ in $\mathcal{G}_4$ so that if U_R belongs to Δ, the Riemann problem may also admit a solution containing three waves with the same zero speed. Let

$$\{U_9\} = \mathcal{W}_1(U_L) \cap \{u = 0\}, \quad U_{10} = \mathcal{W}_1(U_L) \cap \Sigma_3.$$

The solution can be a 1-shock from U_L to some state $E \in \mathcal{G}_3$, followed by a stationary wave using p^* with level $a \in [a_L, a_R]$ that remains in $\mathcal{G}_3$ to some point F, then followed

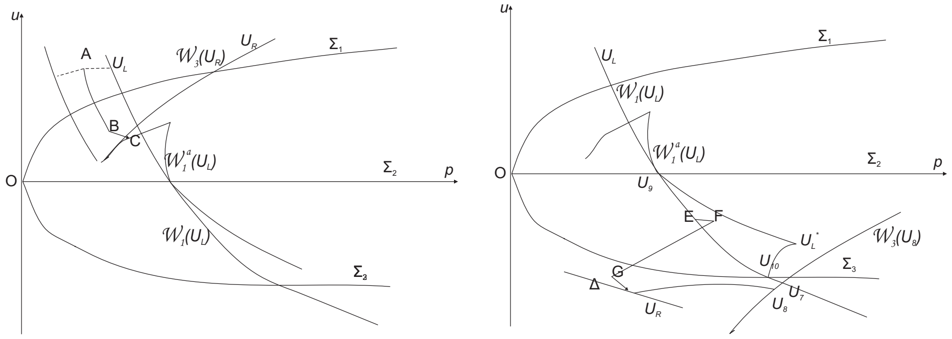


FIG. 4.2. *Riemann solution with structures* (4.7) *(left) and* (4.8) *(right).*

by a 3-shock with zero speed from F to some point $G \in \mathcal{G}_4$, and then followed by a stationary wave from G with level a to U_R with level a_R. For a varies in $[a_L, a_R]$, this procedure forms a set Δ in $\mathcal{G}_4$ so that if $U_R \in \Delta$, then a solution containing three waves with the same zeros speed exists, as just discussed above. See Figure 4.2 (right).

Thus, we can see from Constructions N4 and N5 that multiple solutions may exist.

Existence of Riemann problem. It is interesting to see that the curve $\mathcal{L}(U_L, a_R)$ meets $\mathcal{W}_1(U_1)$ when $a = a_R$, and $\mathcal{L}(U_R, a_R)$ meets $\mathcal{W}_1^a(U_L)$ when $a = a_L$. These patterns form a locally Lipschitz continuous curve in the (p, u)-plane. The Riemann problem admits a solution whenever $\mathcal{W}_3(U_R)$ intersects this curve in Constructions N1–N3, and $U_7 \in \mathcal{G}_4$ in Construction N4. Denote $U_L^* \in \mathcal{G}_3$ to be the state in which a stationary jump from U_{10} is available. Then, a sufficient condition for the Riemann problem to possess a solution is the following:

- U_L^* lies below the curve $\mathcal{W}_3(U_R)$, and U_R is above Σ_3, or $U_R \in \mathcal{G}_4$ and the configuration (4.5) makes sense. In this case, $\mathcal{W}_3(U_R)$ always intersects the curve $\mathcal{W}_1(U_1) \cup \mathcal{L}(U_L, a_R) \cup W_1^a(U_L)$ and the corresponding configuration of solutions makes sense.
- $U_7 \in \mathcal{G}_4$ for U_R is below Σ_3.

Evidently, the Riemann problem admits a solution for a large domain containing U_L.

Uniqueness of Riemann problem. If the right-hand side U_R is chosen so that only Construction N1 makes sense, then we get the *unique* solution. Geometrically, we can see that this can be done if $\mathcal{W}_3(U_R)$ does not meet $\mathcal{W}_1^a(U_L) \cup \mathcal{L}(U_L, a_R)$. Since the mapping $a \mapsto U(a), a \in [a_L, a_R]$ in (4.6) is locally Lipschitz, we can choose $|a_R - a_L|$ not too large so that $\mathcal{L}(U_L, a_R)$ is not far away from $\tilde{U}_L$. An alternative priori estimate for large $|a_R - a_L|$ is that $\mathcal{W}_3(U_R)$ lies entirely in $\mathcal{G}_1$, and this is the case if the intersection of this curve with the axis $\{p = 0\}$ is at a point with nonnegative velocity. Setting $p = 0$ in (2.28), we require

$$u_R - \frac{2\gamma^{1/2}}{(\gamma-1)^{1-1/2\gamma}} \exp\Big(\frac{S_R - S_*}{2C_v\gamma}\Big) p_R^{(\gamma-1)/2\gamma} \geq 0$$

or

$$u_R \geq \frac{2\gamma^{1/2}}{(\gamma-1)^{1-1/2\gamma}} \exp\Big(\frac{S_R - S_*}{2C_v\gamma}\Big) p_R^{(\gamma-1)/2\gamma}. \tag{4.9}$$

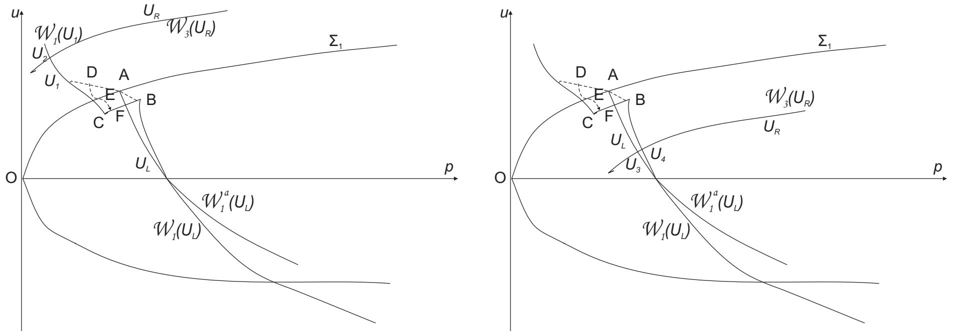


FIG. 4.3. *Riemann solution with structures* (4.10) *(left) and* (4.11)*(right).*

Therefore, we arrive at the following theorem.

THEOREM 4.1 (existence and uniqueness). *Given $U_L \in \mathcal{G}_1$, let $\{U_{10}\} = \mathcal{W}_1(U_L) \cap \Sigma_3$, and let $U_L^* \in \mathcal{G}_3$ be the state resulted by a stationary wave from U_{10} using p^*. The Riemann problem for* (1.1)–(1.2) *always has a solution if U_L^* lies below the curve $\mathcal{W}_3(U_R)$, and U_R is above Σ_3, or $U_R \in \mathcal{G}_4$ and the configuration* (4.5) *makes sense. Moreover, if* (4.9) *holds, or if $|a_R - a_L|$ is sufficiently small, then the Riemann problem for* (1.1)–(1.2) *has* (4.1) *or* (4.2) *as the unique solution.*

4.2. Explicit solutions and uniqueness for $U_L \in \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4 \cup \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$.

Construction N6. This construction holds for U_R belongs in $\mathcal{G}_1 \cup \Sigma_1$ and some part of $\mathcal{G}_2$. The solution begins with a 1-rarefaction wave $R_1(\bar{U}_L, A)$, where $\{A\} = \mathcal{W}_1(U_L) \cap \Sigma_1$; followed by a stationary jump $W_4(A, U_1)$, where $U_1 \in \mathcal{G}_1$, using p^* at A. Let $\{U_2\} = \mathcal{W}_1(U_1) \cap \mathcal{W}_3(U_R)$. The solution is then continued by a 1-wave from U_1 to U_2, followed by a 2-wave $W_2(U_2, \bar{U}_2)$, and finally followed by a 3-wave from $\bar{U}_2$ to U_R. Thus, the solution is

$$R_1(U_L, A) \oplus W_4(A, U_1) \oplus W_1(U_1, U_2) \oplus W_2(U_2, \bar{U}_2) \oplus W_3(\bar{U}_2, U_R). \tag{4.10}$$

The construction makes sense if $\lambda_1(U_1, U_2) \geq 0$. This construction is similar to Construction N1. See Figure 4.3(left).

Construction N7. This construction holds for U_R in $\mathcal{G}_2 \cup G_3 \cup \Sigma_1 \cup \Sigma_2$ and some part of $\mathcal{G}_1$. Let $\{A\} = \mathcal{W}_1(U_L) \cap \Sigma_1$ as in Construction 5, and let $B \in \mathcal{G}_2$ be the point resulted by a stationary wave $W_4(A, B)$ using p^*. Define the "composite" curve

$$\mathcal{W}_1^a(U_L) := \{\bar{U} : \exists W_4(U, \bar{U}) \quad \text{from } a_L \quad \text{to } a_R, U = (p, u) \in \mathcal{W}_1(U_L), p \geq \tilde{p}_A\}. \tag{4.11}$$

Whenever $\mathcal{W}_3(U_R) \cap \mathcal{W}_1^a \neq \emptyset$, there will be a Riemann solution. In fact, let $\mathcal{W}_3(U_R) \cap \mathcal{W}_1^a = \{U_4\}$ and U_3 be the point on $\mathcal{W}_1(U_L)$ that corresponds to the stationary wave $W_4(U_3, U_4)$ or $W_4(\bar{U}_3, U_4)$. Then, the solution can be

$$W_1(U_L, U_3) \oplus W_4(U_3, U_4) \oplus W_2(U_4, \bar{U}_4) \oplus W_3(\bar{U}_4, U_R) \tag{4.12}$$

if $u_3 \geq 0$, and

$$W_1(U_L, U_3) \oplus W_2(U_3, \bar{U}_3) \oplus W_4(\bar{U}_3, U_4) \oplus W_3(U_4, U_R) \tag{4.13}$$

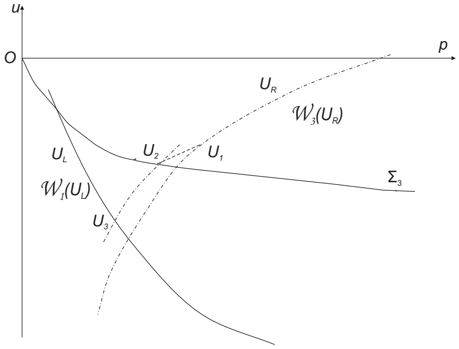


FIG. 4.4. *Riemann solution with structure* (4.14).

if $u_3 < 0$ and $U\lambda_3(U_4, U_R) \geq 0$. This construction is similar to Construction N2. This construction makes sense whenever $U_3 \in \mathcal{G}_2 \cup \mathcal{G}_3 \cup \cup\Sigma_1 \cup \Sigma_2 \cup \Sigma_3$. See Figure 4.3(right).

Construction N8. This construction shows a connection between Constructions N6 and N7. Here, we also meet an interesting phenomenon when wave speeds associated with different characteristic fields coincide; therefore there are solutions containing three waves with the same zero speed. This can be seen as follows. Look at Construction N6; the solution can jump to $\mathcal{G}_2$ as far as $C := \tilde{U}_1$ with a 1-shock with zero speed. Now, instead of jumping from A to U_1 using a $W_4(A, U_1)$ to shift the level a from a_L to a_R, the solution can use a stationary wave from A to some state $D \in \mathcal{G}_1$ with a shift in a from a_L to any value $a \in [a_L, a_R]$, then followed by a 1-shock $S_1(D, E)$ with zero speed, where $E \in \mathcal{G}_2$, and then followed by another stationary wave from E to $F = F(a) \in \mathcal{G}_2$ with a shift in a from $a \in [a_L, a_R]$ to a_R. As a varies continuously on $[a_L, a_R]$, the set of $F(a)$ forms a curve with $F(a_L) = B$ and $F(a_R) = C$, and we have an arc $\widehat{BC} \subset \mathcal{G}_2$. If $G \in \mathcal{W}_3(U_R) \cap \widehat{BC}$, then the Riemann solution admits a solution. This is similar to Construction N3. See Figure 4.3(right).

The arc $\widehat{BC}$ connects the 1-wave curve $\mathcal{W}_1(U_1)$ in Construction N6 and the composite wave curve $\mathcal{W}_1^a(U_L)$ in Construction N7. Therefore, the existence and uniqueness of the Riemann problem can be argued similarly as in the case $U_L \in \mathcal{G}_1$.

Construction N9. This construction holds for $U_L \in \mathcal{G}_4$ and some part of $\mathcal{G}_3 \cup \Sigma_3$; $U_R \in \mathcal{G}_2 \cup \mathcal{G}_3$; see Figure 4.4.

There are $U_1 \in \mathcal{W}_3(U_R)$ and $U_2 \in \Sigma_3$ that correspond to a stationary jump from the left-hand state U_2 with $a = a_L$ to the right-hand state U_1 with $a = a_R$. The state U_1 can be defined in one of the following ways:

(i) We use a backward composite wave curve defined by using a 3-wave from U_R to any $U \in \mathcal{W}_3(U_R)$ followed by a stationary contact from U to some $\hat{U}$ using p^* in the backward way. This curve intersects with Σ_3 at U_2.

(ii) We define a curve going along with Σ_3 by taking all the resulting states $\hat{U}$ of stationary contacts jumping from every point of Σ_3 using p^*. This curve intersects $\mathcal{W}_3(U_R)$ at U_1.

Whenever $\mathcal{W}_1(U_L) \cap \mathcal{W}_3(U_2) \neq \emptyset$, there is a solution defined as follows. Let

$$\{U_3\} = \mathcal{W}_1(U_L) \cap \mathcal{W}_3(U_2).$$

The solution is then

$$W_1(U_L, U_3) \oplus W_2(U_3, \bar{U}_3) \oplus W_3(\bar{U}_3, U_2) \oplus W_4(U_2, U_1) \oplus W_3(U_1, U_R). \tag{4.14}$$

In fact, this construction makes sense for a larger domain of U_R: U_R may belong to $\mathcal{G}_1$ provided the shock speed $\lambda_3(U_1, U_R) \geq 0$.

The above constructions list all possible configurations of Riemann solutions. So we now discuss the existence and uniqueness. Observe that in Construction N7, if $|a_R - a_L|$ is small, then the arc $\widehat{BC}$ is closed to the point A. Therefore, $\mathcal{W}_3(U_R)$ does not meet $\widehat{BC} \cup \mathcal{W}_1(U_1)$ if $|U_R - U_L| + |a_R - a_L|$ is small. Consequently, the Riemann problem admits a unique solution of the form (4.12) or (4.13). In Construction N4, we also have the unique solution if $|U_R - U_L| + |a_R - a_L|$ is small. So we arrive at the following theorem.

THEOREM 4.2 (existence and uniqueness). *Given* $U_L \in \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4 \cup \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$, *let* $\{U_{10}\} = \mathcal{W}_1(U_L) \cap \Sigma_3$, *and let* $U_L^* \in \mathcal{G}_3$ *be the state resulted by a stationary wave from* U_{10} *using* p^*. *The Riemann problem for* (1.1)–(1.2) *always has a solution if* U_L^* *lies below the curve* $\mathcal{W}_3(U_R)$, *and* U_R *is above* Σ_3, *or* $U_R \in \mathcal{G}_4$ *and the configurations* (4.8) *or* (4.13) *make sense. Moreover, we have the following:*

(i) *If* $U_L, U_R \in \mathcal{G}_2 \cup \mathcal{G}_3$ *and* $|U_R - U_L| + |a_R - a_L|$ *is sufficiently small, then the Riemann problem for* (1.1)–(1.2) *has* (4.12) *or* (4.13) *as the unique solution.*

(ii) *If* $U_L, U_R \in \mathcal{G}_4$ *and* $|U_R - U_L| + |a_R - a_L|$ *is sufficiently small, then the Riemann problem for* (1.1)–(1.2) *has* (4.8) *as the unique solution.*

Acknowledgments. The author would like to express his sincere thanks to the referees for their very constructive comments and helpful suggestions.

REFERENCES

[1] R. ABGRALL AND R. SAUREL, *Discrete equations for physical and numerical compressible multiphase mixtures*, J. Comput. Phys., 186 (2003), pp. 361–396.

[2] N. ANDRIANOV AND G. WARNECKE, *On the solution to the Riemann problem for the compressible duct flow*, SIAM J. Appl. Math., 64 (2004), pp. 878–901.

[3] E. AUDUSSE, F. BOUCHUT, M.-O. BRISTEAU, R. KLEIN, AND B. PERTHAME, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comput., 25 (2004), pp. 2050–2065.

[4] M. R. BAER AND J. W. NUNZIATO, *A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials*, Int. J. Multiphase Flows, 12 (1986), pp. 861–889.

[5] R. BOTCHORISHVILI, B. PERTHAME, AND A. VASSEUR, *Equilibrium schemes for scalar conservation laws with stiff sources*, Math. Comp., 72 (2003), pp. 131–157.

[6] R. BOTCHORISHVILI AND O. PIRONNEAU, *Finite volume schemes with equilibrium type discretization of source terms for scalar conservation laws*, J. Comput. Phys., 187 (2003), pp. 391–427.

[7] F. BOUCHUT, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws, and Well-balanced Schemes for Sources*, Front. Math., Birkhäuser Verlag, Basel, 2004.

[8] J. B. BZIL, R. MENIKOFF, S. F. SON, A. K. KAPILA, AND D. S. STEWARD, *Two-phase modeling of a deflagration-to-detonation transition in granular materials: A critical examination of modelling issues*, Phys. Fluids, 11 (1999), pp. 378–402.

[9] G. DAL MASO, P. G. LEFLOCH, AND F. MURAT, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl. (9), 74 (1995), pp. 483–548.

[10] T. N. DINH, R. R. NOURGALIEV, AND T. G. THEOFANOUS, *Understanding the ill-posed two-fluid model,* in Proceedings of the 10th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NERETH-10), Seoul, Korea, 2003.

[11] P. GOATIN AND P. G. LEFLOCH, *The Riemann problem for a class of resonant hyperbolic systems of balance laws*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 881–902.

[12] E. GODLEWSKI AND P.-A. RAVIART, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer-Verlag, New York, 1996.

[13] L. GOSSE, *A well-balanced scheme using nonconservative products designed for hyperbolic systems of conservation laws with source terms*, Math. Models Methods Appl. Sci., 11 (2001), pp. 339–365.

[14] J. M. Greenberg and A. Y. Leroux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[15] J. M. Greenberg, A. Y. Leroux, R. Baraille, and A. Noussair, *Analysis and approximation of conservation laws with source terms*, SIAM J. Numer. Anal., 34 (1997), pp. 1980–2007.

[16] G. Guerra, F. Marcellini, and V. Schleper, *Balance Laws with Integrable Unbounded Sources*, preprint available online at http://arxiv.org/abs/0809.2664v1 (2008).

[17] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.

[18] E. Isaacson and B. Temple, *Convergence of the* 2×2 *Godunov method for a general resonant nonlinear balance law*, SIAM J. Appl. Math., 55 (1995), pp. 625–640.

[19] S. Jin and X. Wen, *Two interface-type numerical methods for computing hyperbolic systems with geometrical source terms having concentrations*, SIAM J. Sci. Comput., 26 (2005), pp. 2079–2101.

[20] S. Jin and X. Wen, *An efficient method for computing hyperbolic systems with geometrical source terms having concentrations*, J. Comput. Math., 22 (2004), pp. 230–249.

[21] B. L. Keyfitz, R. Sander, and M. Sever, *Lack of hyperbolicity in the two-fluid model for two-phase incompressible flow*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 541–563.

[22] D. Kröner, P. G. LeFloch, and M. D. Thanh, *The minimum entropy principle for fluid flows in a nozzle with discontinuous cross-section*, M2AN Math. Model Numer. Anal., 42 (2008), pp. 425–442.

[23] D. Kröner and M. D. Thanh, *Numerical solutions to compressible flows in a nozzle with variable cross-section*, SIAM J. Numer. Anal., 43 (2005), pp. 796–824.

[24] M.-H. Lallemand and R. Saurel, *Pressure Relaxation Procedures for Multiphase Compressible Flows*, INRIA Report, No. 4038, Rocquencourt, France, 2000.

[25] P. G. LeFloch, *Shock Waves for Nonlinear Hyperbolic Systems in Nonconservative Form*, Institute for Mathematics and its Application, Minneapolis, Preprint 593, 1989.

[26] P. G. LeFloch and A. E. Tzavaras, *Representation of weak limits and definition of nonconservative products*, SIAM J. Math. Anal., 30 (1999), pp. 1309–1342.

[27] P. G. LeFloch and M. D. Thanh, *The Riemann problem for fluid flows in a nozzle with discontinuous cross-section*, Commun. Math. Sci., 1 (2003), pp. 763–797.

[28] P. G. LeFloch and M. D. Thanh, *The Riemann problem for the shallow water equations with discontinuous topography*, Commun. Math. Sci., 5 (2007), pp. 865–885.

[29] D. Marchesin and P. J. Paes-Leme, *A Riemann problem in gas dynamics with bifurcation. Hyperbolic partial differential equations*, III, Comput. Math. Appl. Part A, 12 (1986), pp. 433–455.

[30] R. Saurel and R. Abgrall, *A multiphase Godunov method for compressible multifluid and multiphase flows*, J. Comput. Phys., 150 (1999), pp. 425–467.

[31] R. Saurel and R. Abgrall, *A simple method for compressible multifluid flows*, SIAM J. Sci. Comput., 21 (1999), pp. 1115–1145.

[32] M. D. Thanh, Md. Fazlul Karim, and A. Izani Md. Ismail, *Well-balanced scheme for shallow water equations with arbitrary topography*, Int. J. Dyn. Syst. Differ. Equ., 1 (2008), pp. 196–204.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1520–1541

SCATTERING OF SURFACE WATER WAVES BY A FLOATING ELASTIC PLATE IN TWO DIMENSIONS*

RUPANWITA GAYEN† AND B. N. MANDAL‡

Abstract. A new method is developed to study the problem of water wave scattering by a thin elastic plate of arbitrary width floating in deep water assuming linear theory. Using Havelock's expansion of water wave potentials, the boundary value problem describing the potentials is reduced to solving singular integral equations of Carleman type. With the introduction of some integral operators the problem is further reduced to twelve Fredholm integral equations of second kind with regular kernels, and the numerical solutions of these integral equations are used to compute the reflection and transmission coefficients. The numerical estimates for the reflection coefficient are presented in a number of figures given varying different physical parameters. It is shown that the present analysis produces known results for the reflection coefficient.

Key words. water waves, scattering, elastic plate, reflection coefficient

AMS subject classification. 76B

DOI. 10.1137/070685580

1. Introduction.

The problems of water wave scattering by thin elastic plates of either semi-infinite or of finite width have been investigated by a number of researchers using a variety of mathematical techniques. The interest behind investigating this class of problems arises due to their applications in several practical areas. One of these is associated with understanding the behavior of the waves while interacting with the sea ice in the Marginal Ice Zone (MIZ) in Antarctica. Examples of such studies can be found in Fox and Squire (1990), (1994), Meylan and Squire (1994), Squire et al. (1995), and Williams and Squire (2006). The effect of surface wave interaction with floating elastic plates is also important in modeling floating breakwaters and very large floating structures (VLFS) like floating runways, offshore pleasure cities, floating oil-storage bases, etc. Problems dealing with wave-VLFS interaction have been considered by Kagemoto, Masataka, and Motohika (1998), Namba and Okhusu (1999), Kashiwagi (2000), Khapasheva and Korobkin (2002), Okhusu and Namba (2004), and others. An extensive review of ice-wave interaction problems and related methods to solve them are given by Squire (2007).

Evans and Davies (1968) derived the explicit solution of the scattering problem involving a semi-infinite thin elastic plate in finite depth water using the Wiener–Hopf technique (cf. Noble (1958)); however, no numerical calculation could be carried out due to the complicated nature of the solution. Later this problem was also attacked by Balmforth and Craster (1999) and Chung and Fox (2002) using the Wiener–Hopf technique by incorporating some modifications to determine simpler expressions for the reflection coefficient. Gol'dshtein and Marchenko (1989) and Tkacheva (2001), (2003) also employed the Wiener–Hopf technique to study various problems related to floating elastic plates.

*Received by the editors March 18, 2007; accepted for publication (in revised form) November 17, 2008; published electronically March 4, 2009. This work was partially supported by a research grant from DST (SR/S4/MS:263/05)(BNM).

http://www.siam.org/journals/siap/69-6/68558.html

†Department of Mathematics, Indian Institute of Technology, Kharagpur 721302, India (rupanwita@maths.iitkgp.ernet.in).

‡Corresponding author. Physics and Applied Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700 108, India (biren@isical.ac.in).

A variety of different techniques can be found in the water wave literature for studying scattering problems involving two-dimensional models of elastic plates floating on the surface of either finite or infinite depth water. Newman (1994) presented a methodology for treating the interaction of water waves with arbitrary deformable bodies. His idea was to represent the displacement of the bodies in terms of sets of appropriate model functions and orthogonal polynomials. This theory was generalized by Wu, Watanabe, and Utsunomiya (1995) to a single floating elastic plate. Sahoo, Yip, and Chwang (2001) investigated the interaction of surface waves with a semi-infinite elastic plate floating on the surface of finite depth water. They used the method of eigenfunction expansions in the mathematical analysis. A mode matching principle was used by Meylan and Squire (1993) to find the reflection and transmission coefficients of ocean waves by a semi-infinite ice floe (thin elastic plate). In another paper Meylan and Squire (1994) considered a single ice floe of finite width as well as a pair of ice floes of the same width and the related problem was reduced to solving a Fredholm integral equation of second kind with logarithmic kernel by the application of Green's function technique. The problem of scattering of water waves by multiple floating plates of variable properties floating on water of uniform finite depth was considered by Kohout et al. (2007) using the principle of matching of eigenfunction expansions at the boundaries of the plates. Also they compared their solution with experimental results. Andrianov and Hermans (2003) and Hermans (2004) considered a single strip or multiple strips of floating elastic platforms employing integrodifferential equations along the platforms preceded by application of Green's integral theorem. The diffraction of surface waves by a semi-infinite ice sheet modeled as a thin elastic plate and by a gap of finite width between two semi-infinite elastic plates floating on water of finite depth were studied by Linton and Chung (2003) and Chung and Linton (2005), respectively, by the residue calculus technique.

On the other hand, very few papers can be found in the literature involving three-dimensional models of floating elastic plates, as the computations associated with these problems require much rigorous effort compared to two-dimensional problems. Even if the geometry of the plate and the boundary is three-dimensional, for simplicity, some restrictions are imposed on the shape of the plate and/or the boundary to reduce the dimension of the problem to two (cf. Porter and Porter (2004)). A fully developed theory for three-dimensional models can be found in Masson and LeBlond (1989) in connection with wave propagation through rigid circular ice floes in MIZ. They used a multiple scattering theory. Meylan, Squire, and Fox (1997) and Meylan and Masson (2006) studied the wave interaction with flexible ice floes in MIZ for arbitrary floe geometry. Both papers are based on a linear Boltzmann equation formulation. In a recent paper Porter and Evans (2007) considered the three-dimensional problem of scattering of flexural gravity waves through a finite number of cracks of finite length. The method is based on solving hypersingular integral equations by Galerkin technique.

Chakrabarti (2000a) solved explicitly the two-dimensional problem of water wave scattering by a semi-infinite ice-cover floating on the surface of deep water by reducing the problem to solving a Carleman-type singular integral equation. This technique was also employed by Chakrabarti (2000b) to study wave scattering by the discontinuity on the surface of water arising due to the presence of two types of semi-infinite inertial surfaces. A somewhat similar type of problem wherein a single semi-infinite inertial surface is present on the surface of deep water was earlier treated by Kanoria, Mandal, and Chakrabarti (1999) by the Wiener–Hopf technique. They also considered a finite strip of inertial surface floating on the surface of open water.

A Carleman singular integral equation is of the following form (cf. Spence (1965), Varley and Walker (1989)):

$$a(\xi)u(\xi)+\frac{b(\xi)}{\pi}\,⨍_{0}^{\infty}\frac{u(t)}{t-\xi}\mathrm{d}t=c(\xi),\ \ \xi>0, \tag{1.1}$$

where $a(\xi), b(\xi), c(\xi)$ are known rational functions of suitable order and the integral is in the sense of Cauchy principle value. In order to solve (1.1), an appropriate sectionally analytic function in terms of the unknown function $u(\xi)$ is introduced. Then, using Plemelj's formulae, it is reduced to a Riemann–Hilbert problem which is solved in the usual manner (cf. Muskhelishvili (1953), Gakhov (1966)).

Gayen, Mandal, and Chakrabarti (2005), (2006) generalized the problems of Chakrabarti (2000a), (2000b) to solve the problems of water wave scattering by a finite strip of ice-cover and by a strip of inertial surface floating sandwiched between another kind of inertial surface. Both boundary value problems were reduced to two Carleman singular integral equations of the forms

$$a_1(\xi)p_1(\xi)+\frac{1}{\pi}⨍_{0}^{\infty}\frac{p_1(u)}{\xi-u}\mathrm{d}u+\frac{1}{\pi}\int_{0}^{\infty}\frac{p_2(u)\mathrm{e}^{-ul}}{\xi+u}\mathrm{d}u=r_1(\xi), \tag{1.2}$$

$$a_2(\xi)p_2(\xi)+\frac{1}{\pi}⨍_{0}^{\infty}\frac{p_2(u)}{\xi-u}\mathrm{d}u+\frac{1}{\pi}\int_{0}^{\infty}\frac{p_1(u)\mathrm{e}^{-ul}}{\xi+u}\mathrm{d}u=r_2(\xi) \tag{1.3}$$

for determining the unknown functions $p_1(\xi)$ and $p_2(\xi)$, where l is the width of the strip and $a_i(\xi), r_i(\xi)$ $(i=1,2)$ are known functions. Due to the presence of the second integrals in (1.2) and (1.3), it was not possible to reduce the equations to Riemann–Hilbert problems directly, thereby solving the equations explicitly. This difficulty had been overcome by assuming the strip width l to be sufficiently large. Then the third terms on the left-hand sides of (1.2) and (1.3) became exponentially small and the two equations could be solved approximately by reducing them to Riemann–Hilbert problems.

Recently Gayen, Mandal, and Chakrabarti (2007) reinvestigated the finite strip problem considered earlier by Gayen et al. (2006) and solved it for any arbitrary width of strip. The idea was to introduce some integral operators and, with the help of these operators, express the solutions of the Carleman singular integral equations of the forms (1.2) and (1.3) in terms of solutions of four Fredholm integral equations of the second kind, and then compute the solutions numerically.

Keeping in mind the increasing interest in the study of wave interaction problems involving a floating plate due to their immense practical applications, our aim in the present work is to examine the applicability of the aforementioned Carleman singular integral equation method in solving the physical problem of wave scattering by a thin elastic plate floating freely on deep water. The physical problem leads to solving a boundary value problem for a second order partial differential equation involving fifth order partial derivative in the boundary condition. The occurrence of a fifth order derivative makes the problem somewhat difficult to handle compared to the surface strip problem, which involved first order derivative in the boundary condition (cf. Gayen et al. (2007)). As mentioned earlier, this problem was solved by Gayen et al. (2005) for elastic plates of large width. However, for practical purposes the strip width need not always be large. The problem would be more realistic if it could be solved for arbitrary strip widths, as was considered by Meylan and Squire (1994). This motivated us to employ the present analysis. The problem is formulated in terms of reduced velocity potentials in three different fluid regions, two below the semi-infinite free surfaces on either side of the floating plate and one below the finite plate. The conditions of continuity of the velocity potential and velocity across the vertical lines below the end points of the plate followed by Havelock's inversion theorem produce two Carleman-type singular integral equations. We then introduce a singular and a nonsingular integral operator and, after some algebraic manipulations, the entire problem is reduced to solving twelve Fredholm integral equations of the second kind. These are solved numerically by Nystrom's method, and in terms of these solutions the unknown functions that satisfy the Carleman singular integral equations are determined. The reflection and transmission coefficients and the other unknown constants appearing in the boundary value problem are also computed numerically for arbitrary strip width. The numerical estimates of the reflection coefficient ($|R|$) are represented graphically for large as well as moderate strip widths against the wave number. $|R|$ is also depicted against the length of the plate for its different thickness, where the data have been taken to be the same as were used by Meylan and Squire (1994). We obtain almost identical results as those in Meylan and Squire (1994) following a completely different method. Also the results for large width completely agree with those in Gayen et al. (2005). For wider plates, it has been observed that there is an infinite number of frequencies at which complete transmission takes place below the plate, resulting in rapid oscillation in the nature of the curve for $|R|$. This oscillatory nature reduces noticeably with a decrease in the plate width, and when the plate is sufficiently small we cannot find any zero of the reflection coefficient.

In solving problems of different branches of applied mathematics and engineering, Carleman singular integral equations are frequently used. However, to our knowledge no work prior to that of Chakrabarti (2000b) has been done in the field of linear water wave theory using this method. He showed in his two successive papers (Chakrabarti (2000a), (2000b)) how to reduce two particular scattering problems to a single Carleman singular integral equation. We generalized his works in our papers Gayen et al. (2005), (2006) for finite strip problems. Due to the asymptotic nature of the results, the scheme was not applicable for moderate values of strip width. So we were in search of a general method which would work for arbitrary strip widths. We first applied it to a strip of an inertial surface floating between another inertial surface (cf. Gayen et al. (2007)). There we studied the effect of wave propagation by a finite strip of inertial surface lying sandwiched between two other semi-infinite inertial surfaces. The successful implementation of the procedure impelled us to apply it to solve a more complicated boundary value problem involving higher order derivatives in the boundary conditions.

2. Mathematical formulation.

Cartesian coordinates are chosen in which the (x, z)-plane corresponds to the undisturbed upper surface and y-axis pointing vertically downwards. We consider the scattering of a normally incident surface wave train by an elastic plate which occupies the position $y = 0$, $0 \leq x \leq l$, $-\infty < z < \infty$ in deep water. The plate is composed of an elastic material having Young's modulus E and Poisson's ratio ν and is of very small thickness h_0 so that the draft is negligible. Since the plate is infinitely long along the z-direction, we can consider the problem to be two-dimensional in (x, y)-coordinates only. We assume that the motion in the fluid is irrotational, time-harmonic with time dependence $\mathrm{e}^{-\mathrm{i}\sigma t}$, σ being the angular frequency, and that the fluid is inviscid and incompressible. Within the framework of linearized theory, the mathematical problem can be described by a velocity potential $\Phi(x, y; t) = Re\{\phi(x, y)\mathrm{e}^{-\mathrm{i}\sigma t}\}$, where $\phi(x, y)$ is a time-independent complex valued function. It satisfies the Laplace equation

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=0 \quad \text{in the fluid region;} \tag{2.1}$$

the free surface condition

$$K\phi+\phi_y=0 \quad \text{on } y=0,\, (-\infty<x<0)\bigcup(l<x<\infty); \tag{2.2}$$

the plate condition

$$D\frac{\partial^5\phi}{\partial^4 x\partial y}+K\phi+\phi_y=0 \quad \text{on } y=0,\, 0<x<l, \tag{2.3}$$

D being proportional to flexural rigidity of the plate and given by $D=\frac{Eh_0^3}{12(1-\nu^2)\rho g}$, g being the acceleration due to gravity, and $K=\frac{\sigma^2}{g}$; and the bottom condition

$$\nabla\phi\to 0 \quad \text{as } y\to\infty. \tag{2.4}$$

The conditions of no bending moment and no shearing stress at the two ends of the plate are

$$\begin{aligned}&\phi_{xxy}\to 0 \quad \text{as } x\to 0^+, l^- \text{ on } y=0,\\ &\phi_{xxxy}\to 0 \quad \text{as } x\to 0^+, l^- \text{ on } y=0,\end{aligned} \tag{2.5}$$

and the requirements at infinity are

$$\phi\to\begin{cases} \mathrm{e}^{-Ky+\mathrm{i}Kx}+R\mathrm{e}^{-Ky-\mathrm{i}Kx} & \text{as } x\to-\infty,\\ T\mathrm{e}^{-Ky+\mathrm{i}K(x-l)} & \text{as } x\to\infty,\end{cases} \tag{2.6}$$

where R and T are the unknown amplitudes (complex) of the reflected and transmitted waves. Determination of these two quantities is the principal concern here.

It may be noticed that the ice-cover condition given in (2.3) is derived under the assumption that the waves are long compared to the thickness of the ice; i.e., the inertia term is nearly equal to zero. However, if we don't make this assumption, then the boundary condition would have been

$$D\frac{\partial^5\phi}{\partial^4 x\partial y}+(1-\epsilon K)\phi+\phi_y=0 \quad \text{on } y=0,\, 0<x<l.$$

Unless the frequency of the incident wave is very large, for most of the physical problems we can take $(1-\epsilon K)>0$. Now if we divide the above equation by $(1-\epsilon K)$, then it reduces to

$$D'\frac{\partial^5\phi}{\partial^4 x\partial y}+K'\phi+\phi_y=0 \quad \text{on } y=0,\, 0<x<l,$$

where $(D',K')=\frac{(D,K)}{1-\epsilon K}$, which has a form similar to (2.3).

In the next section we reduce the above boundary value problem to two Carleman-type singular integral equations.

3. Reduction to singular integral equations. We first observe that if we solve the Laplace equation (2.1) subject to the plate condition (2.3) by the method of separation of variables, we obtain the following solutions (cf. Chakrabarti, Ahluwalia, and Manam (2003)):

$$\mathrm{e}^{-\lambda Ky\pm \mathrm{i}\lambda Kx},\ \mathrm{e}^{-\lambda_1 Ky\pm \mathrm{i}\lambda_1 Kx},\ \mathrm{e}^{-\bar\lambda_1 Ky\pm \mathrm{i}\bar\lambda_1 Kx},\ \mathrm{e}^{-\lambda_2 Ky\pm \mathrm{i}\lambda_2 Kx},\ \mathrm{e}^{-\bar\lambda_2 Ky\pm \mathrm{i}\bar\lambda_2 Kx}$$

and $\{\xi(D\xi^4+1)\cos\xi y - K\sin\xi y\}\mathrm{e}^{\pm\xi x}$, $\xi\in(0,\infty)$, where λK is the real positive root of the equation

$$Dk^5+k-K=0, \tag{3.1}$$

whose other roots are $(\lambda_1 K,\bar\lambda_1 K),(\lambda_2 K,\bar\lambda_2 K)$ with $\mathrm{Re}(\lambda_1)>0, \mathrm{Re}(\lambda_2)<0, \mathrm{Im}(\lambda_1,\lambda_2)>0$.

Thus in the region below the strip, $\phi(x,y)$ has the form

$$\phi(x,y)=\alpha\mathrm{e}^{-\lambda Ky+\mathrm{i}\lambda Kx}+\beta\mathrm{e}^{-\lambda Ky-\mathrm{i}\lambda K(x-l)}+\chi(x,y),\ 0<x<l. \tag{3.2}$$

Here it may be observed that we have not taken the solutions involving λ_2 and $\bar\lambda_2$ for expressing $\phi(x,y)$ in $0<x<l$, as these do not satisfy the infinite bottom condition (2.4). The first two terms in (3.2) represent the propagating waves with α and β being unknown constants which can be identified with the reflection and transmission coefficients, respectively, through the points $(l,0)$ and $(0,0)$. The function $\chi(x,y)$ is a combination of the solutions $\mathrm{e}^{-\lambda_1 Ky\pm\mathrm{i}\lambda_1 Kx}, \mathrm{e}^{-\bar\lambda_1 Ky\pm\mathrm{i}\bar\lambda_1 Kx}$ and $\{\xi(D\xi^4+1)\cos\xi y-K\sin\xi y\}\mathrm{e}^{\pm\xi x}$, $\xi\in(0,\infty)$. We introduce a reduced potential $\psi(x,y)$ defined by $\phi=\frac{\partial^2\psi}{\partial x^2}$ and express this function in the three regions $x<0$, $0<x<l$, $x>l$ $(y>0)$. The basic reason to set $\phi=\frac{\partial^2\psi}{\partial x^2}$ is to ensure the convergence of the various integrals appearing in the mathematical analysis. Working with ψ ensures avoiding divergent integrals altogether, which is not possible if we work with ϕ.

We now employ Havelock's expansion of water wave potential (cf. Havelock (1929)) to represent $\psi(x,y)$ as

(3.3)
$$\psi(x,y)=-\frac{1}{K^2}\mathrm{e}^{-Ky+\mathrm{i}Kx}-\frac{R}{K^2}\mathrm{e}^{-Ky-\mathrm{i}Kx}+\frac{2}{\pi}\int_0^\infty\frac{A(\xi)}{\xi^2+K^2}L(\xi,y)\mathrm{e}^{\xi x}\mathrm{d}\xi,\ x<0,\ y>0,$$

$$\begin{aligned}\psi(x,y)=&-\frac{1}{\lambda^2K^2}\left\{\alpha\mathrm{e}^{-\lambda Ky+\mathrm{i}\lambda Kx}+\beta\mathrm{e}^{-\lambda Ky-\mathrm{i}\lambda K(x-l)}\right\}\\&-\frac{1}{\lambda_1^2K^2}\left\{A_1\mathrm{e}^{-\lambda_1 Ky+\mathrm{i}\lambda_1 Kx}+A_2\mathrm{e}^{-\lambda_1 Ky-\mathrm{i}\lambda_1 K(x-l)}\right\}\\&-\frac{1}{\bar\lambda_1^2K^2}\left\{A_3\mathrm{e}^{-\bar\lambda_1 Ky+\mathrm{i}\bar\lambda_1 K(x-l)}+A_4\mathrm{e}^{-\bar\lambda_1 Ky-\mathrm{i}\bar\lambda_1 Kx}\right\}\\&+\frac{2}{\pi}\int_0^\infty\frac{B(\xi)\mathrm{e}^{\xi(x-l)}+C(\xi)\mathrm{e}^{-\xi x}}{P(\xi)}M(\xi,y)\mathrm{d}\xi,\ 0<x<l,\ y>0,\end{aligned} \tag{3.4}$$

(3.5)
$$\psi(x,y)=-\frac{T}{K^2}\mathrm{e}^{-Ky+\mathrm{i}K(x-l)}+\frac{2}{\pi}\int_0^\infty\frac{G(\xi)}{\xi^2+K^2}L(\xi,y)\mathrm{e}^{-\xi(x-l)}\mathrm{d}\xi,\ x>l,\ y>0,$$

where A_1, A_2, A_3, A_4 are unknown constants and $A(\xi)$, $B(\xi)$, $C(\xi)$, and $G(\xi)$ are unknown functions of ξ and are such that the integrals in (3.3) to (3.5) are convergent, and

$$L(\xi, y) = \xi \cos \xi y - K \sin \xi y,$$

$$M(\xi, y) = D\xi^5 \cos \xi y + L(\xi, y),$$

and

$$P(\xi) = \xi^2 (D\xi^4 + 1)^2 + K^2.$$

From the representation of the function $\psi(x, y)$ it may be noted that the domain of the variable ξ is $(0, \infty)$, and so whenever ξ appears in the rest of the paper it will be assumed that $\xi > 0$.

Before we proceed further, we would like to state the following theorem, known as Havelock's inversion theorem (cf. Havelock (1929), Ursell (1947)).

THEOREM 3.1. *If a function $H(t)$ defined for $t > 0$ is of class $C^1(0, \infty)$, is absolutely integrable over $(0, \infty)$, and has the integral representation*

$$H(t) = H_0 \mathrm{e}^{-Ky} + \int_0^\infty \widehat{H}(\xi) L(\xi, t) \mathrm{d}\xi, \;\; t > 0,$$

then the constant H_0 and the function $\widehat{H}$ are given by

$$H_0 = 2K \int_0^\infty H(u) \mathrm{e}^{-Ku} \mathrm{d}u$$

and

$$\widehat{H}(\xi) = \frac{2}{\pi} \frac{1}{\xi^2 + K^2} \int_0^\infty H(u) L(\xi, u) \mathrm{d}u.$$

This theorem plays a crucial role in the further development of our method.

After obtaining the expansions of $\psi(x, y)$ given in (3.3)–(3.5) in the open water region and in the plate covered region, we employ the continuity of ψ and $\frac{\partial \psi}{\partial x}$ across the lines $x = 0$ and $x = l$ $(y > 0)$. This gives rise to four relations involving the eight unknown constants and the four unknown functions. To these relations we apply the above theorem and obtain two pairs of representations for the functions $A(\xi)$ and $G(\xi)$ (for details see Gayen et al. (2005)). Elimination of $A(\xi)$ and $G(\xi)$ from their dual representations produces the following Carleman-type singular integral equations for solving the unknown functions $B(\xi)$ and $C(\xi)$:

$$\mu(\xi) B_1(\xi) + \frac{1}{\pi} \fint_0^\infty \frac{B_1(u)}{u - \xi} \mathrm{d}u - \frac{1}{\pi} \int_0^\infty \frac{C_1(u)}{u + \xi} \mathrm{e}^{-ul} \mathrm{d}u = F_B(\xi), \;\; \xi > 0, \tag{3.6}$$

and

$$\mu(\xi) C_1(\xi) + \frac{1}{\pi} \fint_0^\infty \frac{C_1(u)}{u - \xi} \mathrm{d}u - \frac{1}{\pi} \int_0^\infty \frac{B_1(u)}{u + \xi} \mathrm{e}^{-ul} \mathrm{d}u = F_C(\xi), \;\; \xi > 0, \tag{3.7}$$

where

$$(B_1(\xi), C_1(\xi)) = \frac{DK\xi^5}{P(\xi)} \left(B(\xi), C(\xi) \right), \tag{3.8}$$

$$\mu(\xi) = \frac{\xi^2 (D\xi^4 + 1) + K^2}{DK\xi^5} \tag{3.9}$$

and

$$F_B(\xi) = \frac{\lambda-1}{2\lambda^2 K}\left\{\frac{\alpha \mathrm{e}^{\mathrm{i}\lambda Kl}}{\xi-\mathrm{i}\lambda K} + \frac{\beta}{\xi+\mathrm{i}\lambda K}\right\} + \frac{\lambda_1-1}{2\lambda_1^2 K}\left\{\frac{A_1\mathrm{e}^{\mathrm{i}\lambda_1 Kl}}{\xi-\mathrm{i}\lambda_1 K} + \frac{A_2}{\xi+\mathrm{i}\lambda_1 K}\right\}$$

$$+ \frac{\bar{\lambda}_1-1}{2\bar{\lambda}_1^2 K}\left\{\frac{A_3}{\xi-\mathrm{i}\bar{\lambda}_1 K} + \frac{A_4\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}}{\xi+\mathrm{i}\bar{\lambda}_1 K}\right\}, \tag{3.10}$$

$$F_C(\xi) = \frac{\lambda-1}{2\lambda^2 K}\left\{\frac{\alpha}{\xi+\mathrm{i}\lambda K} + \frac{\beta \mathrm{e}^{\mathrm{i}\lambda Kl}}{\xi-\mathrm{i}\lambda K}\right\} + \frac{\lambda_1-1}{2\lambda_1^2 K}\left\{\frac{A_1}{\xi+\mathrm{i}\lambda_1 K} + \frac{A_2\mathrm{e}^{\mathrm{i}\lambda_1 Kl}}{\xi-\mathrm{i}\lambda_1 K}\right\}$$

$$+ \frac{\bar{\lambda}_1-1}{2\bar{\lambda}_1^2 K}\left\{\frac{A_3\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}}{\xi+\mathrm{i}\bar{\lambda}_1 K} + \frac{A_4}{\xi-\mathrm{i}\bar{\lambda}_1 K}\right\}. \tag{3.11}$$

Here it may be mentioned that due to the presence of the second integrals involving $\exp(-ul)$ in (3.6) and (3.7), it is not possible to solve them in a straightforward manner. To get rid of these two terms, Gayen et al. (2005) eliminated them by assuming the strip width to be sufficiently large and, after employing the technique of the Riemann–Hilbert problem, solved the two equations by a sort of iteration process. Here the assumption of largeness of the strip width is not made, and a new method is introduced which is somewhat similar to Gayen et al. (2007). This is explained in section 4.

It will be found in what follows that the solutions of the integral equations (3.6) and (3.7) can be expressed as linear combinations of some known functions multiplied by six unknown constants $\alpha, \beta, A_1, A_2, A_3, A_4$. Once these expressions are obtained they are substituted into the following eight equations for determining the eight unknown constants $R, T, \alpha, \beta, A_1, A_2, A_3, A_4$:

(3.12)
$$\frac{1+R}{2} = \frac{\alpha+\beta\mathrm{e}^{\mathrm{i}\lambda Kl}}{\lambda^2(\lambda+1)} + \frac{A_1+A_2\mathrm{e}^{\mathrm{i}\lambda_1 Kl}}{\lambda_1^2(\lambda_1+1)} + \frac{A_3\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}+A_4}{\bar{\lambda}_1^2(\bar{\lambda}_1+1)} - \frac{2K^3}{\pi}\int_0^\infty \frac{B_1(\xi)\mathrm{e}^{-\xi l}+C_1(\xi)}{\xi^2+K^2}\mathrm{d}\xi,$$

(3.13)
$$\frac{1-R}{2} = \frac{\alpha-\beta\mathrm{e}^{\mathrm{i}\lambda Kl}}{\lambda(\lambda+1)} + \frac{A_1-A_2\mathrm{e}^{\mathrm{i}\lambda_1 Kl}}{\lambda_1(\lambda_1+1)} + \frac{A_3\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}-A_4}{\bar{\lambda}_1(\bar{\lambda}_1+1)} + \frac{2K^2\mathrm{i}}{\pi}\int_0^\infty \frac{B_1(\xi)\mathrm{e}^{-\xi l}-C_1(\xi)}{\xi^2+K^2}\xi\mathrm{d}\xi,$$

(3.14)
$$\frac{T}{2} = \frac{\alpha\mathrm{e}^{\mathrm{i}\lambda Kl}+\beta}{\lambda^2(\lambda+1)} + \frac{A_1\mathrm{e}^{\mathrm{i}\lambda_1 Kl}+A_2}{\lambda_1^2(\lambda_1+1)} + \frac{A_3+A_4\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}}{\bar{\lambda}_1^2(\bar{\lambda}_1+1)} - \frac{2K^3}{\pi}\int_0^\infty \frac{B_1(\xi)+C_1(\xi)\mathrm{e}^{-\xi l}}{\xi^2+K^2}\mathrm{d}\xi,$$

(3.15)
$$\frac{T}{2} = \frac{\alpha\mathrm{e}^{\mathrm{i}\lambda Kl}-\beta}{\lambda(\lambda+1)} + \frac{A_1\mathrm{e}^{\mathrm{i}\lambda_1 Kl}-A_2}{\lambda_1(\lambda_1+1)} + \frac{A_3-A_4\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}}{\bar{\lambda}_1(\bar{\lambda}_1+1)} + \frac{2K^2\mathrm{i}}{\pi}\int_0^\infty \frac{B_1(\xi)-C_1(\xi)\mathrm{e}^{-\xi l}}{\xi^2+K^2}\xi\mathrm{d}\xi,$$

$$(\lambda K)^3(\alpha+\beta\mathrm{e}^{\mathrm{i}\lambda Kl}) + (\lambda_1 K)^3(A_1+A_2\mathrm{e}^{\mathrm{i}\lambda_1 Kl}) + (\bar{\lambda}_1 K)^3(A_3\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}+A_4)$$

$$-\frac{2}{D\pi}\int_0^\infty \{B_1(\xi)\mathrm{e}^{-\xi l}+C_1(\xi)\}\mathrm{d}\xi = 0, \tag{3.16}$$

$$(\lambda K)^4(\alpha-\beta\mathrm{e}^{\mathrm{i}\lambda Kl}) + (\lambda_1 K)^4(A_1-A_2\mathrm{e}^{\mathrm{i}\lambda_1 Kl}) + (\bar{\lambda}_1 K)^4(A_3\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 Kl}-A_4)$$

$$+\frac{2\mathrm{i}}{D\pi}\int_0^\infty \{B_1(\xi)\mathrm{e}^{-\xi l}-C_1(\xi)\}\xi\mathrm{d}\xi = 0, \tag{3.17}$$

$$(\lambda K)^3(\alpha \mathrm{e}^{\mathrm{i}\lambda K l}+\beta)+(\lambda_1 K)^3(A_1\mathrm{e}^{\mathrm{i}\lambda_1 K l}+A_2)+(\bar{\lambda}_1 K)^3(A_3+A_4\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 K l}) \\ -\frac{2}{D\pi}\int_0^\infty \{B_1(\xi)+C_1(\xi)\mathrm{e}^{-\xi l}\}\mathrm{d}\xi=0, \tag{3.18}$$

$$(\lambda K)^4(\alpha \mathrm{e}^{\mathrm{i}\lambda K l}-\beta)+(\lambda_1 K)^4(A_1\mathrm{e}^{\mathrm{i}\lambda_1 K l}-A_2)+(\bar{\lambda}_1 K)^4(A_3-A_4\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 K l}) \\ +\frac{2\mathrm{i}}{D\pi}\int_0^\infty \{B_1(\xi)-C_1(\xi)\mathrm{e}^{-\xi l}\}\xi\mathrm{d}\xi=0. \tag{3.19}$$

The first four equations, (3.12)–(3.15), are obtained by application of Havelock's inversion theorem on the relations derived from the continuity of the functions ψ and $\frac{\partial\psi}{\partial x}$ across the lines $x=0$ and $x=l$ $(y>0)$, whereas (3.16)–(3.19) are the consequence of the conditions (2.4) at the end points of the plate.

4. Solution for arbitrary width of the plate. In this section we solve the two singular integral equations (3.6) and (3.7) for any plate width. For this we first introduce a singular integral operator

$$\mathcal{S}: L^2(0,\infty)\to L^2(0,\infty)$$

and a nonsingular integral operator

$$\mathcal{S}': C^\infty(0,\infty)\to C^\infty(0,\infty)$$

defined by

$$\mathcal{S}f(\xi)=\mu(\xi)f(\xi)+\frac{1}{\pi}\fint_0^\infty \frac{f(u)}{u-\xi}\mathrm{d}u \tag{4.1}$$

and

$$\mathcal{S}'f(\xi)=-\frac{1}{\pi}\int_0^\infty \frac{f(u)\mathrm{e}^{-ul}}{u+\xi}\mathrm{d}u. \tag{4.2}$$

Then (3.6) and (3.7) reduce to the following forms:

$$\mathcal{S}B_1(\xi)+\mathcal{S}'C_1(\xi)=F_B(\xi) \tag{4.3}$$

and

$$\mathcal{S}C_1(\xi)+\mathcal{S}'B_1(\xi)=F_C(\xi). \tag{4.4}$$

It may be observed that the analytical form of the inverse operator $\mathcal{S}^{-1}$ of $\mathcal{S}$ can be determined as follows:

Consider the singular integral equation

$$\mathcal{S}f(\xi)=h(\xi). \tag{4.5}$$

Assuming that the right-hand side is known, (4.5) can be reduced to a Riemann–Hilbert problem (see Muskhelishvili (1953, p. 123), Gakhov (1966, p. 148)),

$$(\mu(\xi)+\mathrm{i})\,\Lambda^+(\xi)-(\mu(\xi)-\mathrm{i})\,\Lambda^-(\xi)=h(\xi), \tag{4.6}$$

after introducing a sectionally analytic function associated with the unknown function $f(\xi)$ satisfying (4.5) as

$$\Lambda(\zeta)=\frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{f(u)}{u-\zeta}\mathrm{d}u, \quad \zeta=\xi+\mathrm{i}\eta. \tag{4.7}$$

$\Lambda(\zeta)$ is defined in the entire complex ζ-plane cut along the real axis from 0 to ∞. One can solve the Riemann–Hilbert problem given in (4.6) by methods of complex variable theory (see England (1971), Estrada and Kanwal (2000)). Plemelj's formulae corresponding to equation (4.7) are

$$\Lambda^{\pm}(\xi) = \pm\frac{1}{2}f(\xi) + \frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{f(u)}{u-\xi}\mathrm{d}u$$

so that

$$\Lambda^{+}(\xi) - \Lambda^{-}(\xi) = f(\xi) \quad \text{and} \quad \Lambda^{+}(\xi) + \Lambda^{-}(\xi) = \frac{1}{\pi\mathrm{i}}\fint_0^\infty \frac{f(u)}{u-\xi}\mathrm{d}u = 2\Lambda(\xi).$$

Thus the function $f(\xi)$ is found to be

$$\begin{aligned} f(\xi) &= \mathcal{S}^{-1}h(\xi) = \Lambda^{+}(\xi) - \Lambda^{-}(\xi) \\ &= \frac{\Lambda_0^{+}(\xi)}{\mu(\xi)-\mathrm{i}}\hat{\mathcal{S}}\left[\frac{h(\xi)}{\Lambda_0^{+}(\xi)(\mu(\xi)+\mathrm{i})}\right], \end{aligned} \tag{4.8}$$

where the operator $\hat{\mathcal{S}}$ is defined by

$$\hat{\mathcal{S}}g(\xi) = \mu(\xi)g(\xi) - \frac{1}{\pi}\fint_0^\infty \frac{g(u)}{u-\xi}\mathrm{d}u \tag{4.9}$$

and

$$\Lambda_0^{+}(\xi) = \lim_{\eta\to 0^+}\Lambda_0(\zeta); \tag{4.10}$$

$\Lambda_0(\zeta)$ is a solution of the homogeneous problem corresponding to the Riemann–Hilbert problem (4.6) and its explicit form is found to be

$$\Lambda_0(\zeta) = \exp\left[\frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{\log\left(\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right) - \lim_{t\to\infty}\log\left(\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right)}{t-\zeta}\mathrm{d}t\right] \quad (\zeta \notin (0,\infty)). \tag{4.11}$$

It may be noted that the limiting term inside the integral is zero. We now apply the operator $\mathcal{S}^{-1}$ to (4.3) to obtain $B_1(\xi)$ in terms of $C_1(\xi)$ as

$$B_1(\xi) = \mathcal{S}^{-1}\left[F_B(\xi) - \mathcal{S}'C_1(\xi)\right] \tag{4.12}$$

and then substitute $B_1(\xi)$ into (4.4). This yields

$$\mathcal{S}C_1(\xi) + \mathcal{S}'\left[\mathcal{S}^{-1}(F_B - \mathcal{S}'C_1)\right](\xi) = F_C(\xi). \tag{4.13}$$

Applying the operator $\mathcal{S}^{-1}$ to the above equation, we find

$$\left[I - \mathcal{L}^2\right]C_1(\xi) = r(\xi), \tag{4.14}$$

where the operator $\mathcal{L} = \mathcal{S}^{-1}\mathcal{S}'$ is noncommutative and its analytical form is determined as

$$\mathcal{L}m(\xi) = -\frac{1}{\pi}\frac{\Lambda_0^{+}(\xi)}{\mu(\xi)-\mathrm{i}}\int_0^\infty \frac{m(u)\mathrm{e}^{-ul}}{(u+\xi)\Lambda_0(-u)}\mathrm{d}u. \tag{4.15}$$

The explicit derivation of the above expression is outlined in the appendix. The right-hand side of (4.14) is

$$r(\xi) = \mathcal{S}^{-1}\left[F_C - S' S^{-1} F_B\right](\xi) \tag{4.16}$$

and can be simplified as

$$r(\xi) = \alpha r_\alpha(\xi) + \beta r_\beta(\xi) + A_1 r_{A_1}(\xi) + A_2 r_{A_2}(\xi) + A_3 r_{A_3}(\xi) + A_4 r_{A_4}(\xi), \tag{4.17}$$

where the functions $r_\alpha(\xi), r_\beta(\xi), r_{A_1}(\xi), r_{A_2}(\xi), r_{A_3}(\xi), r_{A_4}(\xi)$ are given by

$$r_\alpha(\xi) = C_\alpha M(\xi)\left[\frac{1}{\Lambda_0(-\mathrm{i}K\lambda)(\xi + \mathrm{i}K\lambda)} + \frac{\mathrm{e}^{\mathrm{i}\lambda K l}}{\Lambda_0(\mathrm{i}K\lambda)}\int_0^\infty \frac{M_1(\xi,u)}{u - \mathrm{i}K\lambda}\mathrm{d}u\right], \tag{4.18}$$

$$r_\beta(\xi) = C_\beta M(\xi)\left[\frac{\mathrm{e}^{\mathrm{i}\lambda K l}}{\Lambda_0(\mathrm{i}K\lambda)(\xi - \mathrm{i}K\lambda)} + \frac{1}{\Lambda_0(-\mathrm{i}K\lambda)}\int_0^\infty \frac{M_1(\xi,u)}{u + \mathrm{i}K\lambda}\mathrm{d}u\right], \tag{4.19}$$

$$r_{A_1}(\xi) = C_{A_1} M(\xi)\left[\frac{1}{\Lambda_0(-\mathrm{i}K\lambda_1)(\xi + \mathrm{i}K\lambda_1)} + \frac{\mathrm{e}^{\mathrm{i}\lambda_1 K l}}{\Lambda_0(\mathrm{i}K\lambda_1)}\int_0^\infty \frac{M_1(\xi,u)}{u - \mathrm{i}K\lambda_1}\mathrm{d}u\right], \tag{4.20}$$

$$r_{A_2}(\xi) = C_{A_2} M(\xi)\left[\frac{\mathrm{e}^{\mathrm{i}\lambda_1 K l}}{\Lambda_0(\mathrm{i}K\lambda_1)(\xi - \mathrm{i}K\lambda_1)} + \frac{1}{\Lambda_0(-\mathrm{i}K\lambda_1)}\int_0^\infty \frac{M_1(\xi,u)}{u + \mathrm{i}K\lambda_1}\mathrm{d}u\right], \tag{4.21}$$

$$r_{A_3}(\xi) = C_{A_3} M(\xi)\left[\frac{\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 K l}}{\Lambda_0(-\mathrm{i}K\bar{\lambda}_1)(\xi + \mathrm{i}K\bar{\lambda}_1)} + \frac{1}{\Lambda_0(\mathrm{i}K\bar{\lambda}_1)}\int_0^\infty \frac{M_1(\xi,u)}{u - \mathrm{i}K\bar{\lambda}_1}\mathrm{d}u\right], \tag{4.22}$$

$$r_{A_4}(\xi) = C_{A_4} M(\xi)\left[\frac{1}{\Lambda_0(\mathrm{i}K\bar{\lambda}_1)(\xi - \mathrm{i}K\bar{\lambda}_1)} + \frac{\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 K l}}{\Lambda_0(-\mathrm{i}K\bar{\lambda}_1)}\int_0^\infty \frac{M_1(\xi,u)}{u + \mathrm{i}K\bar{\lambda}_1}\mathrm{d}u\right], \tag{4.23}$$

with

$$\begin{gathered} M(\xi) = \frac{\Lambda_0^+(\xi)}{\mu(\xi) - \mathrm{i}}, \quad M_1(\xi,u) = \frac{M(u)\mathrm{e}^{-ul}}{\pi(u+\xi)\Lambda_0(-u)}, \quad C_\alpha = C_\beta = \frac{\lambda - 1}{2\lambda^2 K}, \\ C_{A_1} = C_{A_2} = \frac{\lambda_1 - 1}{2\lambda_1^2 K}, \quad C_{A_3} = C_{A_4} = \frac{\bar{\lambda}_1 - 1}{2\bar{\lambda}_1^2 K}. \end{gathered} \tag{4.24}$$

In order to determine the functions $r_j(\xi)$ given in (4.18)–(4.23) we have utilized the definition of $\mathcal{S}^{-1}$ given in (4.8) together with (4.9). It may be observed that

$$\mathcal{S}^{-1}\left(\frac{1}{\xi + \xi_0}\right) = \frac{M(\xi)}{(\xi + \xi_0)\Lambda_0(-\xi_0)},$$

where ξ_0 is a positive constant.

Equation (4.14) can be regarded as an ordinary integral equation (involving no singular kernel) for solving $C_1(\xi)$. However, since the forcing function $r(\xi)$ is unknown in the sense that it contains the unknown constants α, β, etc., (4.14) cannot be solved directly. In order to overcome this difficulty we introduce two new functions, $U(\xi)$ and $V(\xi)$, in terms of the function $C_1(\xi)$ as

$$[I + \mathcal{L}]C_1(\xi) = U(\xi), \quad [I - \mathcal{L}]C_1(\xi) = V(\xi) \tag{4.25}$$

so that

$$C_1(\xi) = \frac{1}{2}[U(\xi) + V(\xi)] \quad \text{and} \quad \mathcal{L}C_1(\xi) = \frac{1}{2}[U(\xi) - V(\xi)]. \tag{4.26}$$

Then (4.14) can be written in terms of either $U(\xi)$ or $V(\xi)$ as

$$[I-\mathcal{L}]U(\xi)=r(\xi) \tag{4.27}$$

or

$$[I+\mathcal{L}]V(\xi)=r(\xi). \tag{4.28}$$

Because of (4.17) we may express $U(\xi)$ and $V(\xi)$ as

$$U(\xi)=[I-\mathcal{L}]^{-1}r(\xi)=\alpha u_\alpha(\xi)+\beta u_\beta(\xi)+A_1u_{A_1}(\xi)+A_2u_{A_2}(\xi)+A_3u_{A_3}(\xi)+A_4u_{A_4}(\xi) \tag{4.29}$$

and

$$V(\xi)=[I+\mathcal{L}]^{-1}r(\xi)=\alpha v_\alpha(\xi)+\beta v_\beta(\xi)+A_1v_{A_1}(\xi)+A_2v_{A_2}(\xi)+A_3v_{A_3}(\xi)+A_4v_{A_4}(\xi), \tag{4.30}$$

where $u_j(\xi), v_j(\xi)$ (with subscript j denoting $\alpha, \beta, A_1, A_2, A_3, A_4$) are unknown functions. These are determined by incorporating the fact that the integral equation (4.27), along with the relation (4.29), and the integral equation (4.28), along with the relation (4.30), are satisfied simultaneously if $u_j(\xi), v_j(\xi)$ satisfy the following Fredholm integral equations of the second kind:

$$[I-\mathcal{L}]u_j(\xi)=r_j(\xi) \tag{4.31}$$

and

$$[I+\mathcal{L}]v_j(\xi)=r_j(\xi), \tag{4.32}$$

where the subscript j stands for $\alpha, \beta, A_1, A_2, A_3, A_4$. The kernels of the equations and the right-hand sides can be computed from (4.15) and the set of relations (4.18)–(4.23), respectively. We solve the integral equations (4.31) and (4.32) by Nystrom's method to derive the functions $u_j(\xi), v_j(\xi)$ numerically.

In order to solve (4.31) and (4.32), we first need to simplify the forms of the operator $\mathcal{L}$ and the functions $\Lambda_0^+(\xi)$, $\Lambda_0(-\xi)$, and $M(\xi)$. The detailed procedure is given in the appendix.

Now, the solutions $B_1(\xi)$ and $C_1(\xi)$ of the Carleman integral equations (3.6) and (3.7) are determined in a straightforward manner as

$$\begin{aligned} B_1(\xi) &= (\mathcal{S}^{-1}F_B)(\xi)-\mathcal{L}C_1(\xi)=(\mathcal{S}^{-1}F_B)(\xi)-\frac{1}{2}\{U(\xi)-V(\xi)\} \\ &= \alpha B_1^\alpha(\xi)+\beta B_1^\beta(\xi)+A_1B_1^{A_1}(\xi)+A_2B_1^{A_2}(\xi)+A_3B_1^{A_3}(\xi)+A_4B_1^{A_4}(\xi) \end{aligned} \tag{4.33}$$

and

$$\begin{aligned} C_1(\xi) &= \frac{1}{2}\{U(\xi)+V(\xi)\} \\ &= \alpha C_1^\alpha(\xi)+\beta C_1^\beta(\xi)+A_1C_1^{A_1}(\xi)+A_2C_1^{A_2}(\xi)+A_3C_1^{A_3}(\xi)+A_4C_1^{A_4}(\xi). \end{aligned} \tag{4.34}$$

The functions $B_1^j(\xi), C_1^j(\xi)$ (the superscript j having obvious meanings) can be computed from the following relations involving the functions $u_j(\xi)$ and $v_j(\xi)$:

$$B_1^{\alpha}(\xi) = \frac{C_\alpha M(\xi)\mathrm{e}^{\mathrm{i}\lambda K l}}{\Lambda_0(\mathrm{i}K\lambda)(\xi - \mathrm{i}K\lambda)} - \frac{1}{2}\left\{u_\alpha(\xi) - v_\alpha(\xi)\right\},$$

$$B_1^{\beta}(\xi) = \frac{C_\beta M(\xi)}{\Lambda_0(-\mathrm{i}K\lambda)(\xi + \mathrm{i}K\lambda)} - \frac{1}{2}\left\{u_\beta(\xi) - v_\beta(\xi)\right\},$$

$$B_1^{A_1}(\xi) = \frac{C_{A_1} M(\xi)\mathrm{e}^{\mathrm{i}\lambda_1 K l}}{\Lambda_0(\mathrm{i}K\lambda_1)(\xi - \mathrm{i}K\lambda_1)} - \frac{1}{2}\left\{u_{A_1}(\xi) - v_{A_1}(\xi)\right\},$$

$$B_1^{A_2}(\xi) = \frac{C_{A_2} M(\xi)}{\Lambda_0(-\mathrm{i}K\lambda_1)(\xi + \mathrm{i}K\lambda_1)} - \frac{1}{2}\left\{u_{A_2}(\xi) - v_{A_2}(\xi)\right\},$$

$$B_1^{A_3}(\xi) = \frac{C_{A_3} M(\xi)}{\Lambda_0(\mathrm{i}K\bar{\lambda}_1)(\xi - \mathrm{i}K\bar{\lambda}_1)} - \frac{1}{2}\left\{u_{A_3}(\xi) - v_{A_3}(\xi)\right\},$$

$$B_1^{A_4}(\xi) = \frac{C_{A_4} M(\xi)\mathrm{e}^{-\mathrm{i}\bar{\lambda}_1 K l}}{\Lambda_0(-\mathrm{i}K\bar{\lambda}_1)(\xi + \mathrm{i}K\bar{\lambda}_1)} - \frac{1}{2}\left\{u_{A_3}(\xi) - v_{A_3}(\xi)\right\}, \tag{4.35}$$

and

$$C_1^{j}(\xi) = \frac{1}{2}\left\{u_j(\xi) + v_j(\xi)\right\}. \tag{4.36}$$

Thus the functions $B_1(\xi)$ and $C_1(\xi)$ are now derived as linear combinations of unknown constants $\alpha, \beta, A_1, A_2, A_3, A_4$. We then replace $B_1(\xi)$ and $C_1(\xi)$ appearing in (3.12)–(3.19) by their forms in (4.35) and (4.36). This yields a set of eight linear equations for determining the eight unknown constants including R, T. These equations are solved numerically to compute the numerical estimates for the unknown constants. In the next section the numerical results for the reflection coefficient for different parameters are discussed.

5. Numerical results. For numerical computations (except for Figure 1) a characteristic length L proportional to the wavelength is introduced in order to make the different parameters nondimensional. Thus KL, l/L, and D/L^4 represent dimensionless wave number, plate width, and ice-cover parameter, respectively. Because of the energy identity $|R|^2 + |T|^2 = 1$, it is sufficient to present the graphs of $|R|$ only.

In order to establish the correctness of the numerical results obtained by the present analysis, we have compared $|R|$ with the results given in Meylan and Squire (1994). Choosing the same values of various physical quantities such as Young's modulus ($E = 6$ GPa), Poisson's ratio ($\nu = 0.3$), densities of water ($1025\,\mathrm{kgm}^{-3}$) and ice ($922.5\,\mathrm{kgm}^{-3}$), wavelength $100\,\mathrm{m}$, and $g = 9.81\,\mathrm{ms}^{-2}$ as given in Meylan and Squire (1994), $|R|$ is depicted in Figure 1 for different values of thickness of the ice sheet ($h_0 = 1\mathrm{m}$, $2\mathrm{m}$, $5\mathrm{m}$) against its width (floe-length in meters). If Figure 1 is compared with the corresponding figure (Figure 2) of Meylan and Squire (1994), it is obvious that these are almost identical.

Figure 2 shows $|R|$ for plate width $\frac{l}{L} = 10$, ice-cover parameter $\frac{D}{L^4} = 0.001$. Here we notice that when $KL < 0.7$, $|R|$ is almost zero, implying that there occurs total transmission for incident waves with sufficiently smaller frequencies. The continuous line in this figure is drawn on the basis of our present approach, while the triangles represent corresponding data in Gayen, Mandal, and Chakrabarti (2005), wherein the mathematical analysis was based on the assumption of largeness of the plate width. It is evident that the present results completely match the previous ones, which establishes the validity of the theory presented in this paper.

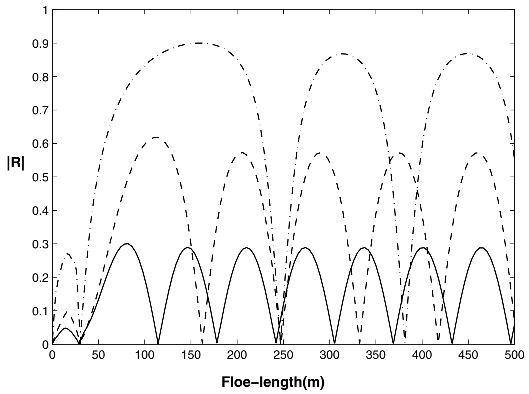


FIG. 1. *$|R|$ for different thickness: $h_0 = 1\mathrm{m}$ (solid curve), $2\mathrm{m}$ (dashed curve), $5\mathrm{m}$ (dash-dot curve).*

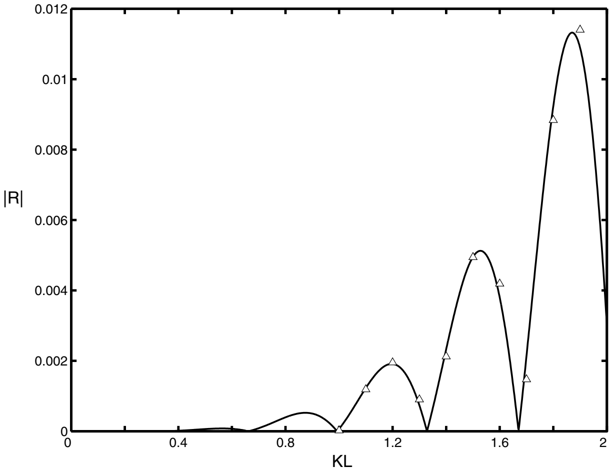


FIG. 2. *$|R|$ for plate width $l/L = 10$, $D/L^4 = 0.001$. Triangles denote data obtained by Gayen, Mandal, and Chakrabarti* (2005). *The line denotes $|R|$ computed by the present method.*

In Figures 3 and 4 the reflection coefficient is plotted against $\frac{D}{L^4}$ for two different wave numbers, $KL = 2$ and $KL = 4$, taking the strip width $\frac{l}{L} = 10$. It is observed that the overall amount of reflection increases with an increase in the wave number. Also the number of zeros of $|R|$ increases for larger frequency.

The effect of frequency is again compared in Figures 5 and 6 by choosing a larger strip width, i.e., $\frac{l}{L} = 100$ for $KL = 2, 4$. From these two figures it is obvious that maximum values of $|R|$ as well as number of zeros of $|R|$ increase with increase of wave number.

Now comparing Figures 3 and 5 or Figures 4 and 6, it is observed that $|R|$ becomes more oscillatory in nature for a wider strip. A similar feature was observed by Chung and Linton (2005) and Williams and Squire (2006) for a gap of finite width between two floating plates where the number of zeros of $|R|$ was found to be greater with larger gap width.

The effect of the thickness of the plate can be examined by varying the parameter

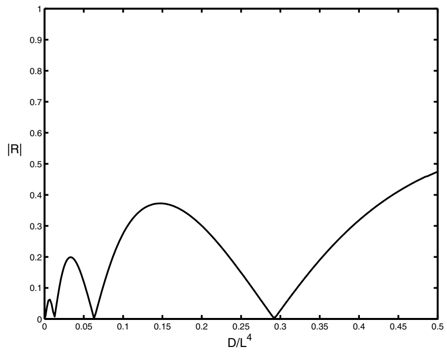


Fig. 3. $|R|$ *for* $l/L = 10,\ KL = 2$.

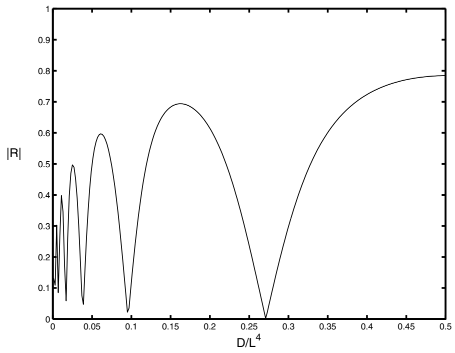


Fig. 4. $|R|$ *for* $l/L = 10,\ KL = 4$.

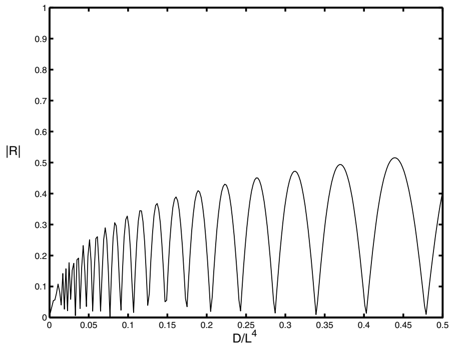


Fig. 5. $|R|$ *for* $l/L = 100,\ KL = 2$.

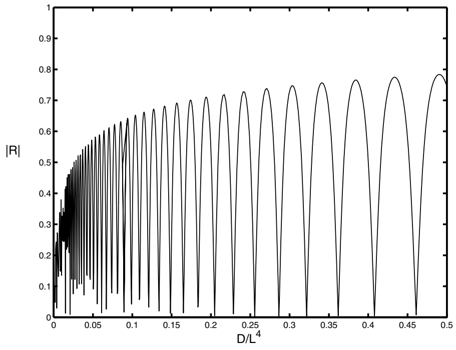


Fig. 6. $|R|$ *for* $l/L = 100,\ KL = 4$.

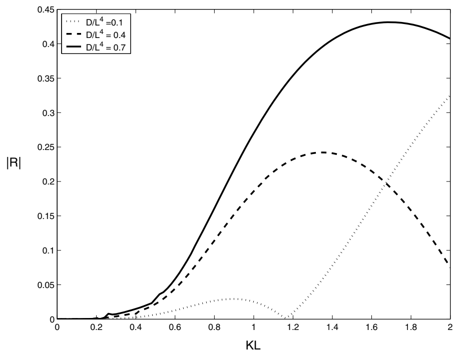


Fig. 7. $|R|$ *for* $l/L = 5$.

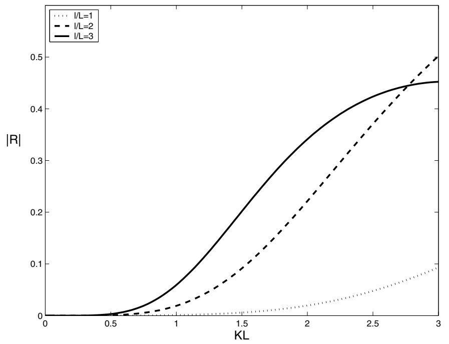


Fig. 8. $|R|$ *for* $D/L^4 = 0.1$.

D if we assume that Young's modulus and Poisson's ratio are kept fixed. This has been shown in Figure 7 by taking $\frac{l}{L} = 5$ and $\frac{D}{L^4} = 0.1, 0.4, 0.7$. The curves in Figure 7 reveal that $|R|$ increases with an increase of $\frac{D}{L^4}$. Thus for plates with the same elastic parameters, the amount of reflected wave energy is increased for thicker plates. This is also evident from Figure 1.

In Figure 8 we have again considered the effect of the plate width on $|R|$ for fixed value 0.1 of D/L^4. Here we have taken moderate plate widths, i.e., $\frac{l}{L} = 1, 2, 3$.

Here also $|R|$ increases with an increase in the plate width, but the phenomenon of multiple reflection does not occur for moderate widths, unlike Figures 3–6, where the plate width was taken to be sufficiently large. It may also be noticed from Figures 7 and 8 that total transmission occurs for smaller frequencies, as was found in Figure 2.

6. Conclusion. In this paper a new method for solving the problem of scattering of a normally incident wave train by the edges of a thin elastic plate of finite width has been described. The boundary value problem is first reduced to singular integral equations of Carleman type and then ultimately to Fredholm integral equations of the second kind. Mathematically it can be claimed that a technique has been developed to study a special type of coupled Carleman singular integral equations of the forms (1.2) and (1.3) wherein the solutions of two complicated singular integral equations have been derived in terms of solutions of integral equations having sufficiently smooth kernels. In this connection, it may be mentioned that Meylan and Squire's (1994) method appears to be somewhat straightforward compared to the method presented here since the solution of a single Fredholm integral equation of the second kind with weakly singular kernel was required. However, the solution method presented here is also somewhat straightforward, and ultimately it involves solving second kind Fredholm integral equations with *regular* kernels. Although our method appears to be cumbersome due to the occurrence of twelve Fredholm integral equations, these equations are solved numerically by using Nystrom's method. A single FORTRAN subroutine served the purpose for solving *all* twelve integral equations. Thus the present method is not really computationally demanding and is perhaps no more difficult than the previous method where a single integral equation is needed to be solved.

A wide range of problems can be handled by the technique presented in this paper. It can be used to investigate wave propagation along strips of floating elastic plates having different thickness or elastic properties, the edges of which are either disjoint (free) or welded together. An identical problem was recently considered by Williams and Squire (2006) using the Wiener–Hopf technique as well as the residue calculus technique. If there are strips of inertial surface surrounded by strips of elastic plates, or vice versa, then our method also can be implemented effectively. This is an ideal situation found in MIZ, where there are continuous sheets of ice and in between them there is broken ice. The former can be modeled as thin elastic plates, whereas the latter can be considered as strips of inertial surface.

The problem of scattering of flexural gravity waves by the edges of two semi-infinite floating elastic plates separated by a finite strip of free surface and present in water of uniform finite depth was investigated by Chung and Linton (2005) employing the residue calculus technique, and the limiting case of this, that is, wave scattering by a narrow crack, was considered by Williams and Squire (2002) and Evans and Porter (2003) using Green's function technique and the mode matching principle. The deep water version of these problems can be studied using the technique of this paper.

One of the classical problems in the literature of water waves is associated with water wave scattering by a semi-infinite or a finite dock (see Friedrichs and Lewy (1948), Holford (1964a), (1964b), Linton (2001), Hermans (2003), Chakrabarti, Mandal, and Gayen (2005)). The effect of wave propagation along a finite dock situated on the surface of deep water with appropriate behaviors at the edges of the dock can be examined by using the present analysis.

Appendix. The operator $\mathcal{L}$. Using the definitions of integral operators $\mathcal{S}'$ and $\mathcal{S}^{-1}$ as given in (4.2), (4.8), and (4.9), we find that

$$\mathcal{L}m(\xi) = (\mathcal{S}^{-1}\mathcal{S}')m(\xi)$$

$$= \frac{\Lambda_0^+(\xi)}{\mu(\xi)-\mathrm{i}}\left[\frac{\mu(\xi)}{\Lambda_0^+(\xi)\,(\mu(\xi)+\mathrm{i})}\left(-\frac{1}{\pi}\int_0^\infty \frac{m(u)\mathrm{e}^{-ul}}{u+\xi}\mathrm{d}u\right)\right.$$

$$\left. + \frac{1}{\pi^2}\int_0^\infty m(u)\mathrm{e}^{-ul}\mathrm{d}u\left(\fint_0^\infty \frac{\mathrm{d}t}{\Lambda_0^+(t)\,(\mu(t)+\mathrm{i})\,(t+u)(t-\xi)}\right)\right]. \tag{A.1}$$

To evaluate the inner integral in the second term of (A.1), we consider the integral

$$\int_\Gamma \frac{\mathrm{d}\tau}{\Lambda_0(\tau)(\tau+u)(\tau-\zeta)}, \quad \zeta \notin \Gamma, \tag{A.2}$$

where $\Lambda_0(\tau)$ satisfies the homogeneous Riemann–Hilbert problem

$$[\mu(\xi)+\mathrm{i}]\,\Lambda^+(\xi) - [\mu(\xi)-\mathrm{i}]\,\Lambda^-(\xi) = 0 \tag{A.3}$$

in the complex τ-plane cut along the positive real axis. The contour Γ is sketched in Figure 9.

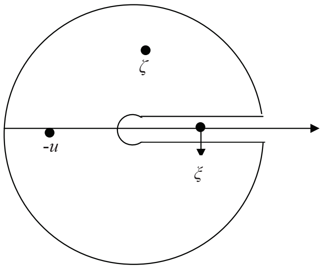


FIG. 9. *The contour* Γ.

Now the integral in (A.2) can be manipulated as

$$\int_\Gamma \frac{\mathrm{d}\tau}{\Lambda_0(\tau)(\tau+u)(\tau-\zeta)} = \int_0^\infty \left(\frac{1}{\Lambda_0^+(t)} - \frac{1}{\Lambda_0^-(t)}\right)\frac{\mathrm{d}t}{(t+u)(t-\zeta)}$$

$$= 2\mathrm{i}\int_0^\infty \frac{\mathrm{d}t}{\Lambda_0^+(t)(\mu(t)+\mathrm{i})(t+u)(t-\zeta)} \tag{A.4}$$

after using (A.3).

Also from the residue calculus theorem,

$$\int_\Gamma \frac{\mathrm{d}\tau}{\Lambda_0(\tau)(\tau+u)(\tau-\zeta)} = \frac{2\pi\mathrm{i}}{u+\zeta}\left(\frac{1}{\Lambda_0(\zeta)} - \frac{1}{\Lambda_0(-u)}\right). \tag{A.5}$$

Comparison of (A.4) and (A.5) gives

$$\frac{1}{u+\zeta}\left(\frac{1}{\Lambda_0(\zeta)} - \frac{1}{\Lambda_0(-u)}\right) = \frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{2\mathrm{i}\mathrm{d}t}{\Lambda_0^+(t)(\mu(\xi)+\mathrm{i})(t+u)(t-\zeta)}. \tag{A.6}$$

Applying Plemelj's formulae to (A.6) the inner integral in the second term on the right-hand side of (A.1) is evaluated as

$$\fint_0^\infty \frac{\mathrm{d}t}{\Lambda_0^+(t)(\mu(t)+\mathrm{i})(t+u)(t-\xi)} = \frac{\pi}{u+\xi}\left(\frac{\mu(\xi)}{(\mu(\xi)+\mathrm{i})\,\Lambda_0^+(\xi)} - \frac{1}{\Lambda_0(-u)}\right).$$

The above relation is substituted into (A.1) and the simplified form of the operator $\mathcal{L}$ is determined as

$$\mathcal{L}m(\xi) = -\frac{1}{\pi}\frac{\Lambda_0^+(\xi)}{\mu(\xi)-\mathrm{i}}\int_0^\infty \frac{m(u)\mathrm{e}^{-ul}}{(u+\xi)\Lambda_0(-u)}\mathrm{d}u.$$

Evaluation of the functions $\boldsymbol{\Lambda_0^+(\xi)}$, $\boldsymbol{\Lambda_0(-\xi)}$, and $\boldsymbol{\mathcal{M}(\xi)}$. We have

$$\Lambda_0(\zeta) = \exp\left[\frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{\log\left[\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right]}{t-\zeta}\mathrm{d}t\right], \quad \zeta \notin (0,\infty), \tag{A.7}$$

where

$$\mu(\xi) \mp \mathrm{i} = \frac{1}{K\xi^5}(\xi \mp \mathrm{i}K)(\xi \pm \mathrm{i}K\lambda)(\xi \pm \mathrm{i}K\lambda_1)(\xi \pm \mathrm{i}K\overline{\lambda}_1)(\xi \pm \mathrm{i}K\lambda_2)(\xi \pm \mathrm{i}K\overline{\lambda}_2). \tag{A.8}$$

If we define

$$\Gamma_0(\zeta) = \log\Lambda_0(\zeta),$$

then

$$\Gamma_0(\zeta) = \frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{\log\left[\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right]}{t-\zeta}\mathrm{d}t \tag{A.9}$$

and

$$\Gamma_0^+(\xi) = \frac{1}{2}\log\left[\frac{\mu(\xi)-\mathrm{i}}{\mu(\xi)+\mathrm{i}}\right] + \frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\left[\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right]}{t-\xi}\mathrm{d}t, \tag{A.10}$$

so that

$$\Lambda_0^+(\xi) = \left[\frac{\mu(\xi)-\mathrm{i}}{\mu(\xi)+\mathrm{i}}\right]^{\frac{1}{2}}\exp[Y(\xi)] \tag{A.11}$$

and

$$M(\xi) = \frac{\Lambda_0^+(\xi)}{\mu(\xi)-\mathrm{i}} = \frac{1}{[1+\mu^2(\xi)]^{\frac{1}{2}}}\exp[Y(\xi)] \tag{A.12}$$

with

$$Y(\xi) = \frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{\log\left[\frac{\mu(t)-\mathrm{i}}{\mu(t)+\mathrm{i}}\right]}{t-\xi}\mathrm{d}t. \tag{A.13}$$

Also it can be shown that

$$\Lambda_0(-u) = \exp[Y(-u)],$$

so that

$$[\Lambda_0(-u)]^{-1}=\exp[-Y(-u)]. \tag{A.14}$$

In the following we proceed to simplify the term $\exp[Y(\xi)]$ only.

Let

$$Y_1(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t-\mathrm{i}K}{t+\mathrm{i}K}}{t-\xi}\mathrm{d}t,\quad Y_2(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t-\mathrm{i}K\lambda}{t+\mathrm{i}K\lambda}}{t-\xi}\mathrm{d}t, \tag{A.15}$$

$$Y_3(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t-\mathrm{i}K\bar{\lambda}_1}{t+\mathrm{i}K\lambda_1}}{t-\xi}\mathrm{d}t,\quad Y_4(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t-\mathrm{i}K\lambda_1}{t+\mathrm{i}K\bar{\lambda}_1}}{t-\xi}\mathrm{d}t,$$

$$Y_5(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t+\mathrm{i}K\lambda_2}{t-\mathrm{i}K\bar{\lambda}_2}}{t-\xi}\mathrm{d}t,\quad Y_6(\xi)=\frac{1}{2\pi\mathrm{i}}\fint_0^\infty \frac{\log\frac{t+\mathrm{i}K\bar{\lambda}_2}{t-\mathrm{i}K\lambda_2}}{t-\xi}\mathrm{d}t.$$

Hence

$$Y(\xi)=Y_1(\xi)-Y_2(\xi)-Y_3(\xi)-Y_4(\xi)+Y_5(\xi)+Y_6(\xi). \tag{A.16}$$

In order to simplify the integrals $Y_j(\xi)$ $(j=1,2,\ldots,6)$, we apply the following result of Varley and Walker (1989):

(A.17)

$$\begin{aligned}V(\xi)&=\frac{1}{2\pi\mathrm{i}}\int_0^\infty \frac{\log\frac{t-\lambda}{t+\lambda}}{t-\xi}\mathrm{d}t\\ &=-\frac{\sin\theta}{\pi}\fint_0^{\frac{\xi}{|\lambda|}}\frac{\ln t}{t^2-2t\cos\theta+1}\mathrm{d}t+\left(1-\frac{\theta}{2\pi}\right)\log\frac{\xi}{\xi-\bar{\lambda}}-\frac{\theta}{2\pi}\log\frac{\xi}{\xi-\lambda},\end{aligned}$$

where $\lambda=|\lambda|\mathrm{e}^{\mathrm{i}\theta}$.

By virtue of the above result, $Y_j(\xi)$'s $(j=1,2,\ldots,6)$ are determined as

$$Y_1(\xi)=-\frac{1}{\pi}\int_0^{\frac{K}{\xi}}\frac{\ln t}{1+t^2}\mathrm{d}t-\mathrm{i}\theta_1+\frac{1}{4}\log\frac{\xi^2}{\xi^2+K^2}, \tag{A.18}$$

$$Y_2(\xi)=-\frac{1}{\pi}\int_0^{\frac{K\lambda}{\xi}}\frac{\ln t}{1+t^2}\mathrm{d}t-\mathrm{i}\theta_2+\frac{1}{4}\log\frac{\xi^2}{\xi^2+\lambda^2K^2}, \tag{A.19}$$

$$\begin{aligned}Y_3(\xi)=&-\frac{\sin\hat{\theta}_3}{\pi}\int_0^{\frac{|\lambda_1|}{\xi}}\frac{\ln t}{t^2-2t\cos\hat{\theta}_3+1}\mathrm{d}t-\frac{\hat{\theta}_3}{2\pi}\log\frac{\xi}{\xi-K(\omega+\mathrm{i}\nu)}\\ &+\left(1-\frac{\hat{\theta}_3}{2\pi}\right)\log\frac{\xi}{\xi-K(\omega-\mathrm{i}\nu)},\end{aligned} \tag{A.20}$$

$$\begin{aligned}Y_4(\xi)=&-\frac{\sin\hat{\theta}_4}{\pi}\int_0^{\frac{|\lambda_1|}{\xi}}\frac{\ln t}{t^2-2t\cos\hat{\theta}_4+1}\mathrm{d}t-\frac{\hat{\theta}_4}{2\pi}\log\frac{\xi}{\xi+K(\omega-\mathrm{i}\nu)}\\ &+\left(1-\frac{\hat{\theta}_4}{2\pi}\right)\log\frac{\xi}{\xi+K(\omega+\mathrm{i}\nu)},\end{aligned} \tag{A.21}$$

$$\begin{aligned}Y_5(\xi)=&-\frac{\sin\hat{\theta}_5}{\pi}\int_0^{\frac{|\lambda_2|}{\xi}}\frac{\ln t}{t^2-2t\cos\hat{\theta}_5+1}\mathrm{d}t-\frac{\hat{\theta}_5}{2\pi}\log\frac{\xi}{\xi-K(\delta+\mathrm{i}\gamma)}\\ &+\left(1-\frac{\hat{\theta}_5}{2\pi}\right)\log\frac{\xi}{\xi-K(\delta-\mathrm{i}\gamma)},\end{aligned} \tag{A.22}$$

$$\begin{aligned} Y_6(\xi) = &-\frac{\sin\hat{\theta}_6}{\pi}\int_0^{\frac{|\lambda_2|}{\xi}}\frac{\ln t}{t^2-2t\cos\hat{\theta}_6+1}\mathrm{d}t-\frac{\hat{\theta}_6}{2\pi}\log\frac{\xi}{\xi+K(\delta-\mathrm{i}\gamma)} \\ &+\left(1-\frac{\hat{\theta}_6}{2\pi}\right)\log\frac{\xi}{\xi+K(\delta+\mathrm{i}\gamma)}, \end{aligned} \tag{A.23}$$

where

$$\begin{aligned} &\theta_1=\tan^{-1}\frac{K}{\xi}, \quad \theta_2=\tan^{-1}\frac{K\lambda}{\xi}, \quad \hat{\theta}_3=\tan^{-1}\frac{\nu}{\omega}, \\ &\hat{\theta}_4=\pi-\hat{\theta}_3, \quad \hat{\theta}_5=\tan^{-1}\frac{\gamma}{\delta}, \quad \hat{\theta}_6=\pi-\hat{\theta}_5, \end{aligned}$$

$\lambda_1=\nu+\mathrm{i}\omega$ and $\lambda_2=-\gamma+\mathrm{i}\delta$; $\nu,\omega,\gamma,\delta>0$.

Substituting the explicit forms of $Y_j(\xi)$ $(j=1,2,\dots,6)$ into (A.16), we ultimately find

$$\begin{aligned} \exp[Y(\xi)] = &\exp[V_{12}(\xi)-V_{34}(\xi)+V_{56}(\xi)]\mathrm{e}^{-\mathrm{i}(\theta_1-\theta_2-\theta_3-\theta_4+\theta_5+\theta_6)} \\ &\times\left|\frac{\xi^2+K^2\lambda^2}{\xi^2+K^2}\right|^{1/4}\left|(\xi-K\omega)^2+K^2\lambda^2\right|^{\frac{1}{2}-\frac{\hat{\theta}_3}{2\pi}}\times\left|(\xi+K\omega)^2+K^2\lambda^2\right|^{\frac{\hat{\theta}_3}{2\pi}} \\ &\times\left|(\xi-K\delta)^2+K^2\gamma^2\right|^{-\frac{1}{2}+\frac{\hat{\theta}_5}{2\pi}}\left|(\xi+K\delta)^2+K^2\gamma^2\right|^{-\frac{\hat{\theta}_5}{2\pi}}, \end{aligned} \tag{A.24}$$

where

$$\begin{aligned} V_{12} &= \frac{1}{\pi}\int_{\frac{K}{\xi}}^{\frac{K\lambda}{\xi}}\frac{\ln t}{1+t^2}\mathrm{d}t, \\ V_{34} &= -\frac{2\sin\hat{\theta}_3}{\pi}\int_0^{\frac{|\lambda_1|}{\xi}}\frac{(t^2+1)\ln t}{(t^2+1)^2-4t^2\cos^2\hat{\theta}_3}\mathrm{d}t, \\ V_{56} &= -\frac{2\sin\hat{\theta}_5}{\pi}\int_0^{\frac{|\lambda_2|}{\xi}}\frac{(t^2+1)\ln t}{(t^2+1)^2-4t^2\cos^2\hat{\theta}_5}\mathrm{d}t, \\ \theta_3,\theta_4 &= \tan^{-1}\frac{K\nu}{\xi\mp K\omega}, \quad \theta_5,\theta_6=\tan^{-1}\frac{K\gamma}{\xi\mp K\delta}. \end{aligned} \tag{A.25}$$

In deriving (A.23) we have used the following results:

$$\xi\mp K\omega+\mathrm{i}K\nu=\left\{(\xi\mp K\omega)^2+K^2\nu^2\right\}^{1/2}\mathrm{e}^{\mathrm{i}(\theta_3,\theta_4)}$$

and

$$\xi\mp K\delta+\mathrm{i}K\gamma=\left\{(\xi\mp K\delta)^2+K^2\gamma^2\right\}^{1/2}\mathrm{e}^{\mathrm{i}(\theta_5,\theta_6)}.$$

Acknowledgments. The authors thank the referees for their comments and suggestions to improve earlier versions of this paper. They also thank Prof. V. A. Squire for providing them with copies of the papers Meylan and Squire (1993), (1994).

REFERENCES

A.I. Andrianov and A.J. Hermans (2003), *The influence of water depth on the hydroelastic response of a very large floating platform*, Marine Structures, 16, pp. 355–371.

N.J. Balmforth and R.V. Craster (1999), *Ocean waves and ice sheets*, J. Fluid Mech., 395, pp. 89–124.

A. Chakrabarti (2000a), *On solution of the problem of scattering of surface water waves by the edge of an ice-cover*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 456, pp. 1087–1099.

A. Chakrabarti (2000b), *On the solution of the problem of scattering of surface water waves by a sharp discontinuity in the surface boundary conditions*, ANZIAM J., 42, pp. 277–286.

A. Chakrabarti, D.S. Ahluwalia, and S.R. Manam (2003), *Surface water waves involving a vertical barrier in the presence of an ice-cover*, Internat. J. Engrg. Sci., 41, pp. 1145–1162.

A. Chakrabarti, B.N. Mandal, and R. Gayen (2005), *The dock problem revisited*, Int. J. Math. Math. Sci., 21, pp. 3459–3470.

H. Chung and C. Fox (2002), *Calculation of wave—ice interaction using the Wiener-Hopf technique*, New Zealand J. Math., 31, pp. 1–18.

H. Chung and C.M. Linton (2005), *Reflection and transmission across a finite gap in an infinite elastic plate on water*, Quart. J. Mech. Appl. Math., 58, pp. 1–15.

A.H. England (1971), *Complex Variable Methods in Elasticity*, Interscience, New York.

R. Estrada and R.P. Kanwal (2000), *Singular Integral Equations*, Birkhäuser, Boston.

D.V. Evans and T.V. Davies (1968), *Water Ice Interaction*, Report 1313, Davidson Laboratory, Stevens Institute of Technology, Hoboken, NJ.

D.V. Evans and R. Porter (2003), *Wave scattering by narrow cracks in ice-sheets floating on water of finite depth*, J. Fluid Mech., 484, pp. 143–165.

C. Fox and V.A. Squire (1990), *Reflection and transmission characteristics at the edge of shore fast sea ice*, J. Geophys. Res., 95, pp. 11625–11639.

C. Fox and V.A. Squire (1994), *On the oblique reflection and transmission of ocean waves at shore fast sea ice*, Philos. Trans. Roy. Soc. London Ser. A, 347, pp. 185–218.

K.O. Friedrichs and H. Lewy (1948), *The dock problem*, Commun. Appl. Math., 1, pp. 135–148.

F.D. Gakhov (1966), *Boundary Value Problems*, Pergamon Press, Oxford.

R. Gayen, B.N. Mandal, and A. Chakrabarti (2005), *Water wave scattering by an ice-strip*, J. Engrg. Math., 53, pp. 21–37.

R. Gayen, B.N. Mandal, and A. Chakrabarti (2006), *Water wave scattering by two sharp discontinuities in the surface boundary conditions*, IMA J. Appl. Math., 71, pp. 811–831.

R. Gayen, B.N. Mandal, and A. Chakrabarti (2007), *Water wave diffraction by a surface strip*, J. Fluid Mech., 571, pp. 419–432.

R.V. Gol'dshtein and A.V. Marchenko (1989), *The diffraction of plane gravitational waves by the edge of an ice-cover*, J. Appl. Math. Mech., 53, pp. 731–736.

T.H. Havelock (1929), *Forced surface waves on water*, Phil. Mag., 8, pp. 569–576.

A.J. Hermans (2003), *Interaction of free surface waves with a floating dock*, J. Engrg. Math., 45, pp. 39–53.

A.J. Hermans (2004), *Interaction of free surface waves with floating flexible strips*, J. Engrg. Math., 49, pp. 133–147.

K.L. Holford (1964a), *Short surface waves in the presence of a finite dock* I, Proc. Cambridge Philos. Soc., 60, pp. 957–983.

K.L. Holford (1964b), *Short surface waves in the presence of a finite dock* II, Proc. Cambridge Philos. Soc., 60, pp. 985–1011.

H. Kagemoto, F. Masataka, and M. Motohika (1998), *Theoretical and experimental predictions of the hydroelastic response of a very large floating structure in waves*, Appl. Ocean Res., 20, pp. 135–144.

M. Kanoria, B.N. Mandal, and A. Chakrabarti (1999), *The Wiener-Hopf solution of a class of mixed boundary value problems arising in surface water wave phenomena*, Wave Motion, 29, pp. 267–292.

M. Kashiwagi (2000), *Research on hydroelastic responses of VLFS: Recent progress and further work*, Int. J. Offshore and Polar Engrg., 10, pp. 81–90.

T.I. Khapasheva and A.A. Korobkin (2002), *Hydroelastic behavior of compound floating plate in waves*, J. Engrg. Math., 44, pp. 21–40.

A.L. Kohout, M.H. Meylan, S. Sakai, K. Hanai, P. Leman, and D. Brossard (2007), *Linear water wave propagation through multiple elastic plates of variable properties*, J. Fluids Struct., 23, pp. 649–663.

C.M. Linton (2001), *The finite dock problem*, Z. Angew. Math. Phys., 52, pp. 640–656.

C.M. Linton and H. Chung (2003), *Reflection and transmission at the ocean/sea-ice boundary*, Wave Motion, 38, pp. 43–52.

D. Masson and P.H. LeBlond (1989), *Spectral evolution of wind- generated surface gravity waves in a dispersed ice field*, J. Fluid Mech., 202, pp. 43–81.

M.H. Meylan and D. Masson (2006), *LAN, a linear Boltzmann equation to model wave scattering in the marginal ice zone*, Ocean Model., 11, pp. 417–427.

M.H. Meylan and V.A. Squire (1993), *Finite floe reflection and transmission coefficients from a semi-infinite model*, J. Geophys. Res., 98(C7), pp. 12537–12542.

M.H. Meylan and V.A. Squire (1994), *The response of ice floes to ocean waves*, J. Geophys. Res., 99(C1), pp. 891–900.

M.H. Meylan, V.A. Squire, and C. Fox (1997), *Toward realism in modelling ocean wave behaviour in marginal ice zones*, J. Geophys. Res., 102 (C10), pp. 22981–22991.

N.I. Muskhelishvili (1953), *Singular Integral Equations*, Noordhoff, Groningen, Holland.

Y. Namba and M. Okhusu (1999), *Hydroelastic behavior of artificial islands in waves*, Int. J. Offshore and Polar Engrg., 9, pp. 39–47.

J.N. Newman (1994), *Wave effect on deformable bodies*, Appl. Ocean Res., 16, pp. 47–59.

B. Noble (1958), *Methods Based on the Wiener–Hopf Technique*, Pergamon Press, New York.

M. Okhusu and Y. Namba (2004), *Hydroelastic analysis of a large floating structure*, J. Fluids Struct., 19, pp. 543–555.

D. Porter and R. Porter (2004), *Approximations to wave scattering by an ice sheet of variable thickness over undulating bed topography*, J. Fluid Mech., 509, pp. 145–179.

R. Porter and D.V. Evans (2007), *Diffraction of flexural waves by finite straight cracks in an elastic sheet over water*, J. Fluids Struct., 23, pp. 309–327.

T. Sahoo, T.L. Yip, and A.T. Chwang (2001), *Scattering of surface waves by a semi-infinite floating elastic plate*, Phys. Fluids, 13, pp. 3215–3222.

D.A. Spence (1965), *Flow past a thin wing with an oscillating jet flap*, Philos. Trans. Roy. Soc. London Ser. A, 257, pp. 445–467.

V.A. Squire (2007), *Of ocean waves and sea-ice revisited*, Cold Regions Sci. Technol., 49, pp. 110–133.

V.A. Squire, J.P. Dugan, P. Wadahams, P.J. Rottier, and A.K. Liu (1995), *Of ocean waves and ice-sheets*, Annu. Rev. Fluid Mech., 27, pp. 115–168.

L.A. Tkacheva (2001), *Scattering of surface waves by the edge of a floating elastic plate*, J. Appl. Mech. Tech. Phys., 42, pp. 638–646.

L.A. Tkacheva (2003), *Plane problem of surface wave diffraction on a floating elastic plate*, Fluid Dynam., 38, pp. 465–481.

F. Ursell (1947), *The effect of a fixed vertical barrier on surface water waves in deep water*, Proc. Cambridge Philos. Soc., 42, pp. 374–382.

E. Varley and J.D.A. Walker (1989), *A method for solving singular integro-differential equations*, IMA J. Appl. Math., 43, pp. 11–45.

T.D. Williams and V.A. Squire (2002), *Wave propagation across an oblique crack in an ice-sheet*, Int. J. Offshore and Polar Engrg., 12, pp. 157–162.

T.D. Williams and V.A. Squire (2006), *Scattering of flexural-gravity waves at the boundaries between three floating sheets with applications*, J. Fluid Mech., 569, pp. 113–140.

C. Wu, E. Watanabe, and T. Utsunomiya (1995), *An eigenfunction expansion matching method for analysing the wave-induced responses of an elastic floating plate*, Appl. Ocean Res., 17, pp. 301–310.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1542–1558

INTERACTIONS OF ELEMENTARY WAVES FOR THE AW–RASCLE MODEL*

MEINA SUN†

Abstract. In this paper, we study the interactions of elementary waves for the traffic flow model proposed by Aw and Rascle in [*SIAM J. Appl. Math.*, 60 (2000), pp. 916–938]. The solutions are obtained constructively when the initial data are three piecewise constant states. In particular, a new wave SJ in which a shock wave S and a contact discontinuity J coincide with each other is obtained during the process of interaction. Moreover, by studying the limits of the solutions as the perturbed parameter ε tends to zero, it can be found that the Riemann solutions are stable for such perturbations with the initial data.

Key words. interaction of elementary waves, vacuum state, Aw–Rascle model, Riemann problem, traffic flow, hyperbolic conservation laws

AMS subject classifications. 35L65, 35L67, 35B30

DOI. 10.1137/080731402

1. Introduction.

The Aw–Rascle (AR) macroscopic model of traffic flow in the conservative form [2] is given by

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho(u + p(\rho)))_t + (\rho u(u + p(\rho)))_x = 0, \end{cases} \tag{1.1}$$

where ρ, u represent the density and the velocity, respectively; the velocity offset p takes the form $p(\rho) = \rho^\gamma$ with $\gamma > 0$. The AR model describes a traffic flow model on a unidirectional roadway. The basic assumptions of the model are the density $\rho(x,t) \geq 0$ and velocity $u(x,t) \geq 0$ of cars located at position x at time t.

The AR model was proposed in order to remedy the deficiencies of second order models of car traffic pointed out by Daganzo [6] and has been independently derived by Zhang [20]. The derivation of the model from a microscopic follow-the-leader (FL) model through a scaling limit was also given in [1]. The AR model resolves all the obvious inconsistencies and explains instabilities in the car traffic flow, especially near the vacuum, i.e., for very light traffic with few slow drivers [2, 11].

The AR model is one of the main fluid dynamic models for traffic flow and is appropriate for describing traffic phenomena, such as congestion and stop-and-go waves [10]. It is now widely used to study the formation and dynamics of traffic jams and is endowed with desirable stability properties. It is also the basis for the multilane traffic flow model [8, 9], the model for a road network with unidirectional flow [7, 10], and the hybrid traffic flow model [13].

In [3], the limit behavior was investigated by changing p into εp and taking $p(\rho) = (\frac{1}{\rho} - \frac{1}{\rho^*})^{-\gamma}$ with the density constraint $\rho \leq \rho^*$, where the maximal density ρ^* corresponds to a total traffic jam and is assumed to a fixed constant although

*Received by the editors July 28, 2008; accepted for publication (in revised form) November 17, 2008; published electronically March 4, 2009. This work was supported by National Natural Science Foundation of China.

http://www.siam.org/journals/siap/69-6/73140.html

†School of Mathematics and Information, Ludong University, Yantai, Shandong 264025, People's Republic of China, and Laboratory of Mathematics Physics, Wuhan Institute of Physics and Mathematics, The Chinese Academy of Sciences, Wuhan 430071, People's Republic of China (smnwhy0350@163.com).

it should depend on the velocity in practice. They discovered that the pressure term becomes active so as to preserve the constraint $\rho \leq \rho^*$ when ρ reaches ρ^*. Recently, Shen and Sun [15] have considered the limit behavior without the constraint of the maximal density, i.e., $p(\rho)$ is not singular at $\rho = \rho^*$. The delta-shock wave was obtained through perturbing the pressure $p(\rho)$ suitably.

For convenience and conciseness, we replace $\rho p(\rho)$ with $p(\rho)$ in (1.1) and take $p(\rho) = \rho^\gamma$ for $\gamma > 1$; then the AR model can be rewritten in the following form:

$$(1.2) \qquad \begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u + \rho^\gamma)_t + (\rho u^2 + \rho^\gamma u)_x = 0. \end{cases}$$

In the above equations, $p(\rho) = \rho^\gamma$ can be regarded as the traffic pressure term and γ is analogous with the adiabatic gas constant in gas dynamics.

In this paper, our main purpose is to investigate various possible interactions of elementary waves for the AR model (1.2). To include all kinds of interactions, it suffices to consider the AR model (1.2) with the following perturbed initial data:

$$(1.3) \qquad (u,\rho)(x,0) = \begin{cases} (u_-, \rho_-), & -\infty < x < -\varepsilon, \\ (u_m, \rho_m), & -\varepsilon < x < \varepsilon, \\ (u_+, \rho_+), & \varepsilon < x < +\infty, \end{cases}$$

where $\varepsilon > 0$ is arbitrarily small. We notice that (1.3) is a local perturbation of the Riemann data and we still call it a small perturbation here for ε is sufficiently small.

Aw and Rascle have investigated the Riemann problem of (1.1) in detail. With these results in mind, one would naturally like to study the interactions of elementary waves because they embody the internal mechanism of the AR model. Another motivation of this study comes from the fact that small changes in traffic flow will propagate and lead to the occurrence of wave interaction. Finally, the stability of the Riemann solutions of (1.2) can be analyzed if we take the initial data (1.3) and then let $\varepsilon \to 0$.

By definition, a vacuum state is any portion of the (x,t) plane in which $\rho = 0$. From [2], we know that the Riemann solutions do involve the vacuum state for certain Riemann data. In order to cover all the cases completely, we divide our work into two parts according to the presence of vacuum or not. When the vacuum is not involved, the problem about the interactions of elementary waves is classical and will not be addressed here. On the other hand, the AR model suitably explains instabilities near the vacuum. Therefore, we especially pay attention to the vacuum problem and consider the interactions of elementary waves in full detail when the vacuum is involved. Dealing with the vacuum problem, we adopt the idea proposed by Liu and Smoller [12] when they considered it for the isentropic gas dynamic equations, where they made a distinction between two vacuum states with different (fake) velocities.

With the method of characteristic analysis, the interactions are widely investigated and the global solutions are completely constructed. Furthermore, we find that the solutions of the perturbed initial value problem (1.2) and (1.3) converge to the solutions of the corresponding Riemann problem (1.2) and (2.1) as $\varepsilon \to 0$, which shows the stability of the Riemann solutions for certain perturbations of the initial data. Especially, when the vacuum is involved, the interesting feature in the solutions is that a new wave SJ is discovered during the interaction of a contact discontinuity J and a shock S in a particular situation. Here SJ is the superposition of a contact discontinuity and a shock. The reason for the generation of the wave SJ is due to the fact that the newly formed waves S and J after interaction propagate with the same speed and coincide with each other.

For basic references on nonlinear hyperbolic systems of conservation laws and the interactions of elementary waves, we refer the readers to the book of Smoller [17] and the monograph of Chang and Hsiao [4]. Furthermore, one can see the books written by Dafermos [5] and Serre [14] for a comprehensive survey. Also see [16, 18] for the recent work about the interactions of elementary waves.

This paper is organized as follows. In section 2, we restate the Riemann problem to the AR model (1.2) for readers' convenience. In section 3, we mainly discuss the interactions of elementary waves when the vacuum is involved. In section 4, we consider the stability of the Riemann solutions under the small perturbations and compare our results with those of Aw and Rascle before our conclusion in section 5.

2. Preliminaries. In this section, we briefly review the Riemann solutions of (1.2) with the initial data

$$(u,\rho)(x,0)=(u_\pm,\rho_\pm),\qquad \pm x>0, \tag{2.1}$$

where $u_\pm, \rho_\pm > 0$, and the detailed study can be found in [2].

The characteristic roots of system (1.2) are

$$\lambda_1=u-(\gamma-1)\rho^{\gamma-1},\qquad \lambda_2=u; \tag{2.2}$$

therefore (1.2) is strictly hyperbolic except for $\rho = 0$.

The corresponding right characteristic vector of $\lambda_i (i = 1, 2)$ is

$$\overrightarrow{r_1}=((1-\gamma)\rho^{\gamma-2},1)^T,\qquad \overrightarrow{r_2}=(0,1)^T. \tag{2.3}$$

It is easy to see that $\nabla\lambda_1\cdot\overrightarrow{r_1}\neq 0$ for $\rho\neq 0$ and $\nabla\lambda_2\cdot\overrightarrow{r_2}\equiv 0$ in which ∇ denotes the gradient with respect to (u,ρ); namely, λ_1 is genuinely nonlinear for $\rho\neq 0$ and λ_2 is always linearly degenerate. Therefore, the associated waves are rarefaction waves or shocks for the first family and contact discontinuities for the second family.

The Riemann invariants along the characteristic fields are

$$w=u+\rho^{\gamma-1},\qquad z=u. \tag{2.4}$$

Since (1.2) and the Riemann data (2.1) are invariant under stretching of coordinates: $(x,t)\to(\alpha x,\alpha t)$ (α is constant), we seek the self-similar solution

$$(u,\rho)(x,t)=(u,\rho)(\xi),\qquad \xi=x/t. \tag{2.5}$$

Then the Riemann problem is reduced to the boundary value problem of the ordinary differential equations:

$$\begin{cases} -\xi\rho_\xi+(\rho u)_\xi=0,\\ -\xi(\rho u+\rho^\gamma)_\xi+(\rho u^2+\rho^\gamma u)_\xi=0, \end{cases} \tag{2.6}$$

with $(u,\rho)(\pm\infty)=(u_\pm,\rho_\pm)$.

For smooth solutions, setting $U=(u,\rho)^T$, (2.6) can then be rewritten as

$$A(U)U_\xi=0, \tag{2.7}$$

where

$$A(u,\rho) = \begin{pmatrix} \rho & u-\xi \\ -\xi\rho + 2\rho u + \rho^\gamma & -\xi u - \gamma\xi\rho^{\gamma-1} + u^2 + \gamma\rho^{\gamma-1}u \end{pmatrix}.$$

Besides the constant state solution, it provides a rarefaction wave which is a continuous solution of (2.7) in the form $(u,\rho)(\xi)$. Then, for a given left state (u_-,ρ_-), the rarefaction wave curves in the phase plane, which are the sets of states that can be connected on the right by a 1-rarefaction wave, are as follows:

$$R(u_-,\rho_-) : \begin{cases} \xi = \lambda_1 = u - (\gamma-1)\rho^{\gamma-1}, \\ u - u_- = -\rho^{\gamma-1} + \rho_-^{\gamma-1}, \\ \rho < \rho_-, \qquad u > u_-. \end{cases} \tag{2.8}$$

Through differentiating u with respect to ρ in the second equation in (2.8), we get

$$u_\rho = -(\gamma-1)\rho^{\gamma-2}, \qquad u_{\rho\rho} = -(\gamma-1)(\gamma-2)\rho^{\gamma-3}. \tag{2.9}$$

Thus the 1-rarefaction wave curve is convex for $1 < \gamma < 2$ and concave for $\gamma > 2$ in the (u,ρ) plane.

For a bounded discontinuity at $\xi = \sigma$, the Rankine–Hugoniot condition holds:

$$\begin{cases} -\sigma[\rho] + [\rho u] = 0, \\ -\sigma[\rho u + \rho^\gamma] + [\rho u^2 + \rho^\gamma u] = 0, \end{cases} \tag{2.10}$$

where $[\rho] = \rho_r - \rho_l$, $\rho_l = \rho(\sigma - 0)$, and $\rho_r = \rho(\sigma + 0)$, etc.

From the first equation in (2.10), we obtain

$$\rho_r(u_r - \sigma) = \rho_l(u_l - \sigma). \tag{2.11}$$

Simplifying the second equation in (2.10) and noting (2.11), it yields

$$\rho_r(u_r - \sigma)(u_r + \rho_r^{\gamma-1} - u_l - \rho_l^{\gamma-1}) = 0. \tag{2.12}$$

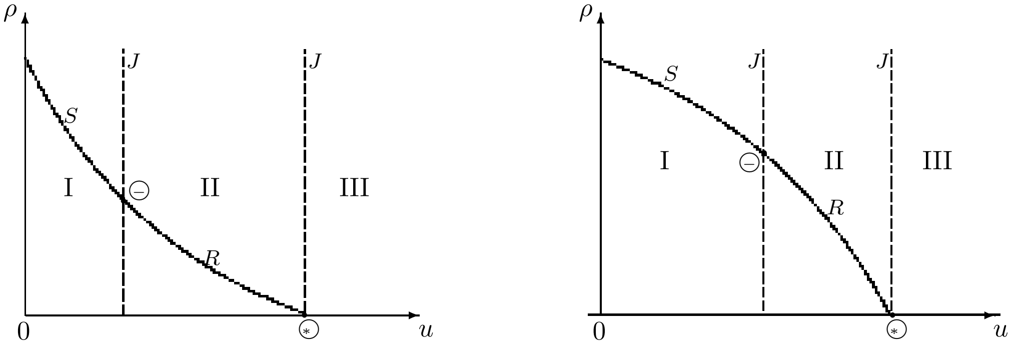
If $\rho_r(u_r - \sigma) \neq 0$, we have $u_r + \rho_r^{\gamma-1} = u_l + \rho_l^{\gamma-1}$, and the Lax entropy conditions imply that $\rho_l < \rho_r$. So for a given left state (u_-,ρ_-), the sets of states which can be connected to (u_-,ρ_-) by a 1-shock wave on the right are as follows:

$$S(u_-,\rho_-) : \begin{cases} \sigma = u - \dfrac{\rho_-(\rho^{\gamma-1} - \rho_-^{\gamma-1})}{\rho - \rho_-}, \\ u - u_- = -\rho^{\gamma-1} + \rho_-^{\gamma-1}, \\ \rho > \rho_-, \qquad u < u_-. \end{cases} \tag{2.13}$$

It is noted that the shock curves coincide with the rarefaction curves in the phase plane, due to the special form of (1.2), which can be written as $Y_t + (uY)_x = 0$ [19].

If $\rho_r(u_r - \sigma) = 0$, we can conclude that $u_r = u_l = \sigma$ except for $\rho_r = 0$ or $\rho_l = 0$, which corresponds to a contact discontinuity of the second family. Since λ_2 is linearly degenerate, the sets of states can be connected to a given left state (u_-,ρ_-) by a contact discontinuity on the right if and only if

$$J : \xi = u = u_-. \tag{2.14}$$



$1 < \gamma < 2$ $\gamma > 2$

FIG. 2.1.

We note that these waves travel at exactly the same speed as the corresponding cars, which means no information travels faster than the vehicle velocity and the drivers do not react to the traffic situation behind him.

Through the above analysis, we summarize that the sets of states connected on the right consist of the 1-rarefaction wave curve $R(u_-, \rho_-)$, the 1-shock wave curve $S(u_-, \rho_-)$, and the 2-contact discontinuity $J(u_-, \rho_-)$ for a given left state (u_-, ρ_-). These curves divide the quarter phase plane $(u, \rho \geq 0)$ into three regions, $\mathrm{I} = \{(u, \rho)|u < u_-\}$, $\mathrm{II} = \{(u, \rho)|u_- < u < u_*\}$, and $\mathrm{III} = \{(u, \rho)|u > u_*\}$, where $u_* = u_- + \rho_-^{\gamma-1}$ (see Figure 2.1). According to the right state (u_+, ρ_+) in the different region, one can construct the unique global Riemann solution connecting two constant states $(u_\pm, \rho_\pm)$.

Obviously, the Riemann solution contains a 1-shock wave, an intermediate nonvacuum constant state, and a 2-contact discontinuity when $(u_+, \rho_+) \in \mathrm{I}$; it contains a 1-rarefaction wave, an intermediate nonvacuum constant state, and a 2-contact discontinuity when $(u_+, \rho_+) \in \mathrm{II}$; it contains a 1-rarefaction wave, an intermediate vacuum state, and a 2-contact discontinuity when $(u_+, \rho_+) \in \mathrm{III}$.

All of the rarefaction waves R, the shock waves S, and the contact discontinuities J obtained in solving the Riemann problem are called the elementary waves for the AR model (1.2).

3. Interactions of elementary waves. We begin by considering the initial value problem (1.2) with three pieces of constant initial data (1.3). The data (1.3) is a perturbation of the Riemann initial data (2.1). We face the interesting question of determining whether the Riemann solutions of (1.2) and (2.1) are the limits of $(u_\varepsilon, \rho_\varepsilon)(x, t)$ as $\varepsilon \to 0$, where $(u_\varepsilon, \rho_\varepsilon)(x, t)$ are the solutions of (1.2) and (1.3). We will deal with this problem case by case along with constructing the solutions.

We notice that the Riemann solutions of (1.2) and (2.1) may contain the vacuum, so in order to cover all the cases, our discussion should be divided into two parts according to the appearance of vacuum or not. About the interactions of elementary waves not involving the vacuum, we have four cases according to the different combinations of elementary waves from $(-\varepsilon, 0)$ and $(\varepsilon, 0)$ as follows:

1. $R + J$ and $S + J$,
2. $S + J$ and $S + J$,
3. $S + J$ and $R + J$,
4. $R + J$ and $R + J$.

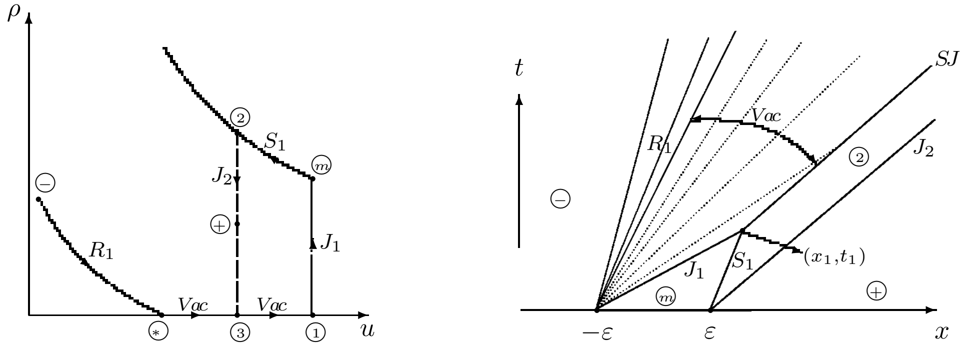


FIG. 3.1.

All the above interactions are classical and well known; hence they will not be pursued here.

In this section, we mainly consider the interactions of elementary waves when at least one of the Riemann solutions at $(-\varepsilon, 0)$ and $(\varepsilon, 0)$ involves the vacuum state. When the Riemann solution contains the vacuum, the AR model (1.2) becomes degenerate in the vacuum region and the two characteristics coincide. In this work, we can study this problem in the (u, ρ) plane, i.e., make a distinction between two vacuum states with different (fake) velocities, like for the method introduced by Liu and Smoller [12] for compressible gas dynamics.

Also, our discussion is divided into the following five cases according to the different combinations of elementary waves from $(-\varepsilon, 0)$ and $(\varepsilon, 0)$:

1. $R + Vac + J$ and $S + J$,
2. $R + Vac + J$ and $R + J$,
3. $R + Vac + J$ and $R + Vac + J$,
4. $R + J$ and $R + Vac + J$,
5. $S + J$ and $R + Vac + J$.

Case 3.1. $R + Vac + J$ and $S + J$.

In this case, when t is small, the solution of the initial value problem (1.2) and (1.3) can be expressed briefly as follows (see Figures 3.1 and 3.2):

$$(u_-, \rho_-) + R_1 + Vac + J_1 + (u_m, \rho_m) + S_1 + (u_2, \rho_2) + J_2 + (u_+, \rho_+),$$

where "+" means "followed by." This case happens if and only if $u_\pm < u_m$ and $u_* = u_- + \rho_-^{\gamma-1} < u_m$.

In the following figures, we just depict the convex situation, i.e., $1 < \gamma < 2$, for the reason that the concave situation is similar.

The propagating speed of J_1 is $\tau_1 = u_m$, and the propagating speed of S_1 is given by

$$\sigma_1 = u_m - \frac{\rho_2(\rho_2^{\gamma-1} - \rho_m^{\gamma-1})}{\rho_2 - \rho_m};$$

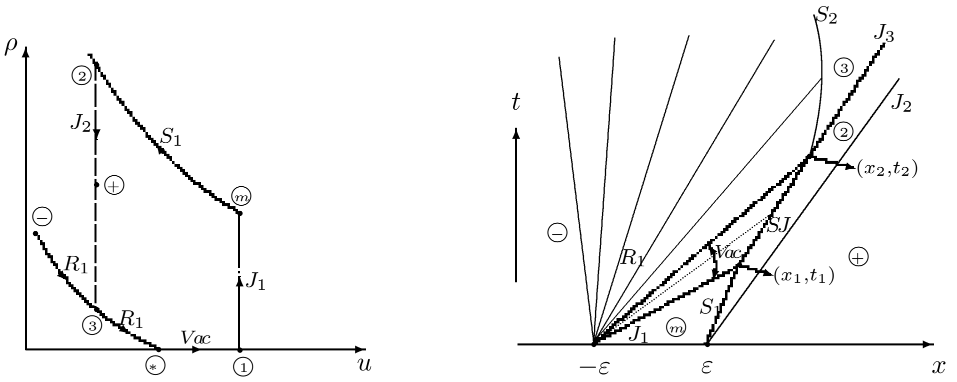


FIG. 3.2.

thus $\tau_1 > \sigma_1$ and the contact discontinuity J_1 will overtake the shock wave S_1 in finite time. The intersection (x_1, t_1) is determined by

$$(3.1)\qquad \begin{cases} x_1 + \varepsilon = u_m t_1, \\ x_1 - \varepsilon = \left(u_m - \dfrac{\rho_2(\rho_2^{\gamma-1} - \rho_m^{\gamma-1})}{\rho_2 - \rho_m} \right) t_1. \end{cases}$$

An easy calculation leads to

$$(3.2)\qquad (x_1, t_1) = \left(\frac{2\varepsilon(\rho_2 - \rho_m)u_m}{\rho_2(\rho_2^{\gamma-1} - \rho_m^{\gamma-1})} - \varepsilon, \frac{2\varepsilon(\rho_2 - \rho_m)}{\rho_2(\rho_2^{\gamma-1} - \rho_m^{\gamma-1})} \right).$$

It is clear that two elementary waves intersect at a finite time when a new Riemann problem is formed. At the time $t = t_1$, we again have a Riemann problem with data $(u_l, \rho_l) = (u_1, \rho_1)$, $(u_r, \rho_r) = (u_2, \rho_2)$, which is resolved by a new shock S and a new contact discontinuity J. Here we notice that at the left-hand side of S is the vacuum state; thus it is not difficult to see that the propagating speeds of S and J are both equal to u_2; i.e., S and J coincide with each other and form a new wave, which we denote by SJ.

Now, we turn our attention to the interaction of SJ and R_1. It is easy to see that $u_2 = u_+$ and the wave front in R_1 propagates with speed u_*. Our claim is that R_1 and SJ cannot intersect if $u_* \leq u_+$ (see Figure 3.1), while for $u_* > u_+$ they must intersect with each other (see Figure 3.2).

If $u_* > u_+$, SJ will cancel the vacuum region and then intersect with R_1 at the point (x_2, t_2), which can be given by

$$(3.3)\qquad \begin{cases} x_2 + \varepsilon = u_* t_2 = (u_- + \rho_-^{\gamma-1})t_2, \\ x_2 - x_1 = u_2(t_2 - t_1). \end{cases}$$

Solving the Riemann problem at (x_2, t_2), we can see the appearance of a shock wave S_2 and a contact discontinuity J_3. Namely, when $t > t_2$, SJ decomposes and the state (u_3, ρ_3) lies between S_2 and J_3. At the same time, the shock wave S_2 begins to penetrate R_1 with a varying speed of propagation during the process of penetration;

that is, the shock $S_2 : x = x(t)$ is no longer a straight line at $t > t_2$. The varying speed of S_2 can be determined by

$$\begin{cases} \dfrac{dx}{dt} = u - \dfrac{\rho_3(\rho_3^{\gamma-1} - \rho^{\gamma-1})}{\rho_3 - \rho}, \\ x + \varepsilon = (u - (\gamma-1)\rho^{\gamma-1})t, \\ u - u_- = \rho_-^{\gamma-1} - \rho^{\gamma-1}, \\ x(t_2) = x_2, \quad 0 \le \rho < \rho_3. \end{cases} \tag{3.4}$$

Differentiating the second equation in (3.4) with respect to t, we obtain

$$\frac{dx}{dt} = u - (\gamma-1)\rho^{\gamma-1} + t\Big(\frac{du}{dt} - (\gamma-1)^2\rho^{\gamma-2}\frac{d\rho}{dt}\Big). \tag{3.5}$$

Combining (3.5) with the first equation in (3.4), it is easy to get

$$\frac{(\gamma-1)\rho^\gamma - \gamma\rho_3\rho^{\gamma-1} + \rho_3^\gamma}{\rho - \rho_3} = t\Big(\frac{du}{dt} - (\gamma-1)^2\rho^{\gamma-2}\frac{d\rho}{dt}\Big). \tag{3.6}$$

It follows from the third equation in (3.4) that

$$\frac{du}{dt} = -(\gamma-1)\rho^{\gamma-2}\frac{d\rho}{dt}. \tag{3.7}$$

Substituting (3.7) into (3.6), it yields

$$\frac{d\rho}{dt} = \frac{(\gamma-1)\rho^\gamma - \gamma\rho_3\rho^{\gamma-1} + \rho_3^\gamma}{(\gamma-1)\gamma\rho^{\gamma-2}(\rho_3 - \rho)t}. \tag{3.8}$$

Differentiating the first equation in (3.4), in view of (3.7), (3.8), we have

$$\frac{d^2x}{dt^2} = \frac{((\gamma-1)\rho^\gamma - \gamma\rho_3\rho^{\gamma-1} + \rho_3^\gamma)^2}{(\gamma-1)\gamma\rho^{\gamma-2}(\rho - \rho_3)^3 t}, \tag{3.9}$$

which gives $\frac{d^2x}{dt^2} < 0$ for $\rho < \rho_3$, i.e., S_2 decelerates during the process of penetration.

Integrating (3.8) leads to

$$\ln\frac{t}{t_2} = \int_0^\rho \frac{(\gamma-1)\gamma\rho^{\gamma-2}(\rho_3 - \rho)}{(\gamma-1)\rho^\gamma - \gamma\rho_3\rho^{\gamma-1} + \rho_3^\gamma}d\rho. \tag{3.10}$$

It is clear that $t \to \infty$ as $\rho \to \rho_3$. Therefore, S_2 cannot penetrate over R_1 forever if $\rho_3 \le \rho_-$; otherwise S_2 will cross the whole of R_1 at the finite time

$$t_3 = t_2 \exp\left(\int_0^{\rho_-} \frac{(\gamma-1)\gamma\rho^{\gamma-2}(\rho_3 - \rho)}{(\gamma-1)\rho^\gamma - \gamma\rho_3\rho^{\gamma-1} + \rho_3^\gamma}d\rho\right).$$

Thus we conclude that S_2 is able to cross the whole of R_1 for $\rho_3 > \rho_-$ (i.e., $u_+ < u_-$), whereas it cannot for $\rho_3 \le \rho_-$ (i.e., $u_+ \ge u_-$) and ultimately has $x + \varepsilon = (u_3 - (\gamma-1)\rho_3^{\gamma-1})t$ as its asymptote.

In brief, if $u_+ \ge u_*$, when $t > t_1$, the solution can be expressed as

$$(u_-, \rho_-) + R_1 + Vac + SJ + (u_2, \rho_2) + J_2 + (u_+, \rho_+).$$

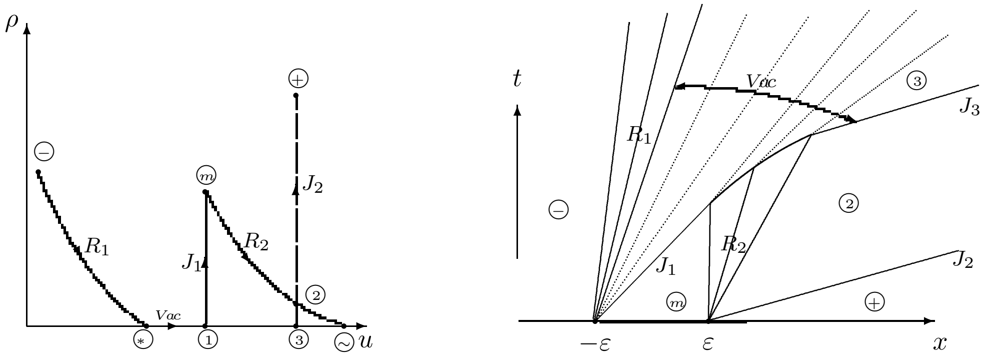


FIG. 3.3.

If $u_+ < u_*$, when $t \to \infty$, the solution can be expressed as

$$(u_-, \rho_-) + R + (u_3, \rho_3) + J_3 + (u_2, \rho_2) + J_2 + (u_+, \rho_+) \quad \text{for } u_+ \geq u_-,$$

$$(u_-, \rho_-) + S + (u_3, \rho_3) + J_3 + (u_2, \rho_2) + J_2 + (u_+, \rho_+) \quad \text{for } u_+ < u_-.$$

Case 3.2. $R + Vac + J$ and $R + J$.

In this case, when t is small, the solution of the initial value problem (1.2) and (1.3) can be expressed briefly as follows (see Figure 3.3):

$$(u_-, \rho_-) + R_1 + Vac + J_1 + (u_m, \rho_m) + R_2 + (u_2, \rho_2) + J_2 + (u_+, \rho_+).$$

This case occurs when $u_* < u_m < u_+ < u_\sim = u_m + \rho_m^{\gamma-1}$ is satisfied.

Obviously, the contact discontinuity J_1 will overtake the rarefaction wave R_2, and they begin to interact with each other at (x_1, t_1), which satisfies

$$\begin{cases} x_1 + \varepsilon = u_m t_1, \\ x_1 - \varepsilon = (u_m - (\gamma - 1)\rho_m^{\gamma-1}) t_1. \end{cases} \tag{3.11}$$

This gives

$$(x_1, t_1) = \left(\frac{2\varepsilon u_m - \varepsilon(\gamma-1)\rho_m^{\gamma-1}}{(\gamma-1)\rho_m^{\gamma-1}}, \frac{2\varepsilon}{(\gamma-1)\rho_m^{\gamma-1}} \right). \tag{3.12}$$

Then J_1 goes on to penetrate R_2, and the contact discontinuity $x = x(t)$ during the process of penetration is determined by

$$\begin{cases} \dfrac{dx}{dt} = u, \\ x - \varepsilon = (u - (\gamma - 1)\rho^{\gamma-1}) t, \\ u - u_m = \rho_m^{\gamma-1} - \rho^{\gamma-1}, \\ x(t_1) = x_1, \quad \rho_2 \leq \rho \leq \rho_m. \end{cases} \tag{3.13}$$

Differentiating (3.13) with respect to t along $x = x(t)$, we obtain

$$\frac{d^2 x}{dt^2} = \frac{du}{dt}, \tag{3.14}$$

$$\frac{dx}{dt} = u - (\gamma-1)\rho^{\gamma-1} + \Big(\frac{du}{dt} - (\gamma-1)^2\rho^{\gamma-2}\frac{d\rho}{dt}\Big)t, \tag{3.15}$$

$$\frac{du}{dt} = -(\gamma-1)\rho^{\gamma-2}\frac{d\rho}{dt}. \tag{3.16}$$

Substituting $\frac{dx}{dt} = u$ into the above expressions, it yields

$$\frac{d^2x}{dt^2} = \frac{(\gamma-1)\rho^{\gamma-1}}{\gamma t} > 0,$$

which means that the contact discontinuity accelerates during the process of penetration.

It follows from (3.13) that

$$\frac{dx}{dt} = \frac{x-\varepsilon}{t} + (\gamma-1)\rho^{\gamma-1} = \frac{x-\varepsilon}{t} + (\gamma-1)\Big(\rho_m^{\gamma-1} + u_m - \frac{dx}{dt}\Big).$$

So (3.13) can be simplified as

$$\begin{cases} \dfrac{dx}{dt} = \dfrac{x-\varepsilon}{\gamma t} + \dfrac{\gamma-1}{\gamma}(u_m + \rho_m^{\gamma-1}), \\ x(t_1) = x_1. \end{cases} \tag{3.17}$$

By applying the method of variation of constant, we obtain

$$x = \varepsilon + (u_m + \rho_m^{\gamma-1})t - \gamma\Big(\frac{2\varepsilon\rho_m}{\gamma-1}\Big)^{1-\frac{1}{\gamma}} t^{\frac{1}{\gamma}}, \tag{3.18}$$

which, together with

$$x - \varepsilon = (u_2 - (\gamma-1)\rho_2^{\gamma-1})t, \tag{3.19}$$

determines the ending point (x_2, t_2) of the penetration. A direct calculation leads to

$$(x_2, t_2) = \Big(\varepsilon + \frac{2\varepsilon\rho_m u_2}{(\gamma-1)\rho_2^{\gamma}} - \frac{2\varepsilon\rho_m}{\rho_2}, \frac{2\varepsilon\rho_m}{(\gamma-1)\rho_2^{\gamma}}\Big). \tag{3.20}$$

It turns out that the contact discontinuity J_1 crosses the rarefaction wave R_2 completely in finite time and R_2 becomes the vacuum state after penetration.

For large time, the solution can be expressed as

$$(u_-, \rho_-) + R_1 + Vac + J_3 + (u_2, \rho_2) + J_2 + (u_+, \rho_+).$$

Case 3.3. $R + Vac + J$ and $R + Vac + J$.

In this case, when t is small, the solutions of the initial value problems (1.2) and (1.3) can be expressed briefly as follows (see Figure 3.4):

$$(u_-, \rho_-) + R_1 + Vac + J_1 + (u_m, \rho_m) + R_2 + Vac + J_2 + (u_+, \rho_+).$$

The occurrence of this case depends on the condition $u_* < u_m < u_\sim < u_+$.

Indeed, this case is similar to Case 3.2 except that the vacuum states appear in front of R_2 at the beginning. In the same way as before, we can see that the

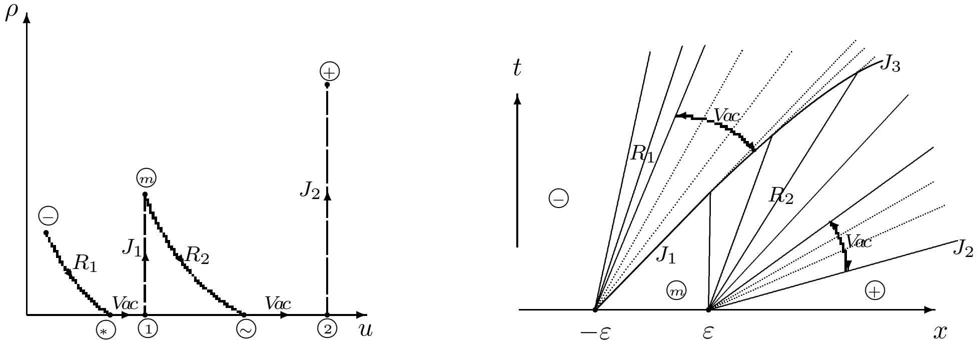


FIG. 3.4.

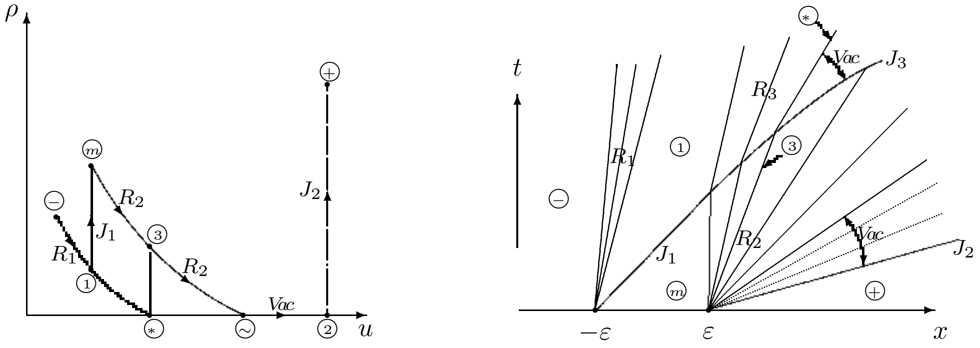


FIG. 3.5.

propagating speed of J_3 tends to $u_\sim$ as $t \to \infty$; i.e., J_3 has the wave front in R_2 as its asymptote.

As $t \to \infty$, the time-asymptotic solution can be described as

$$(u_-, \rho_-) + R_1 + Vac + J_2 + (u_+, \rho_+).$$

Case 3.4. $R + J$ and $R + Vac + J$.

In this case, when t is small, the solutions of the initial value problems (1.2) and (1.3) can be expressed briefly as follows (see Figure 3.5):

$$(u_-, \rho_-) + R_1 + (u_1, \rho_1) + J_1 + (u_m, \rho_m) + R_2 + Vac + J_2 + (u_+, \rho_+).$$

This case happens when $u_- < u_m < u_*$ and $u_+ > u_\sim$ are satisfied.

This case can be discussed similarly to Case 3.3. Moreover, it can be shown that the vacuum states form ahead of R_3 at the time when one of the states in R_2 becomes (u_3, ρ_3). Then, J_3 continues to penetrate R_2 and finally disappears in the vacuum as $t \to \infty$.

For large time, the solution can be expressed as

$$(u_-, \rho_-) + R_1 + (u_1, \rho_1) + R_3 + Vac + J_2 + (u_+, \rho_+).$$

Case 3.5. $S + J$ and $R + Vac + J$.

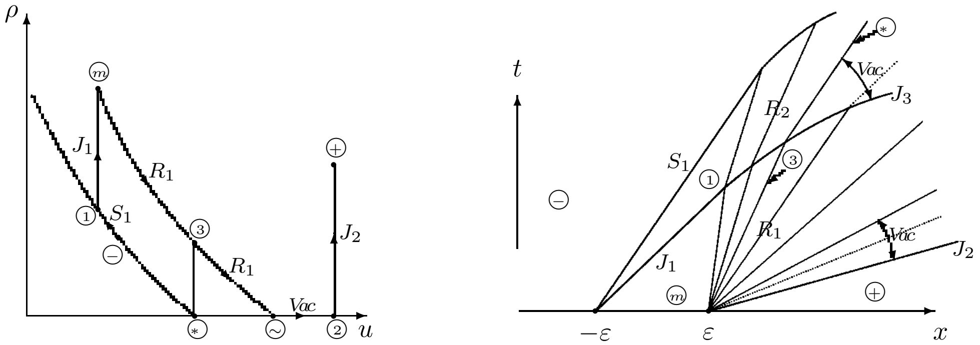


FIG. 3.6.

In this case, when t is small, the solutions of the initial value problems (1.2) and (1.3) can be expressed briefly as follows (see Figure 3.6):

$$(u_-, \rho_-) + S_1 + (u_1, \rho_1) + J_1 + (u_m, \rho_m) + R_1 + Vac + J_2 + (u_+, \rho_+).$$

The appearance of this case depends on the conditions $u_m < u_-$ and $u_\sim < u_+$.

Like for Case 3.4, the interaction of J_1 and R_1 results in a new contact discontinuity J_3 and a new rarefaction wave R_2. Similarly, the vacuum states present ahead of R_2 at the time when one of the states in R_1 turns to be (u_3, ρ_3). The propagating speed of J_3 tends to $u_\sim$ as $t \to \infty$; i.e., J_3 has the wave front in R_1 as its asymptote.

Now we mainly consider the interaction of the shock wave S_1 and the rarefaction wave R_2. The propagating speed of S_1 is

$$\sigma_1 = u_- - \frac{\rho_1(\rho_1^{\gamma-1} - \rho_-^{\gamma-1})}{\rho_1 - \rho_-},$$

and the propagating speed of the wave back in R_2 is $\omega_2 = u_1 - (\gamma-1)\rho_1^{\gamma-1}$. Noting $u_1 - u_- = \rho_-^{\gamma-1} - \rho_1^{\gamma-1}$, it is easy to get

$$\sigma_1 - \omega_2 = \frac{(\gamma-1)\rho_1^\gamma - \gamma\rho_-\rho_1^{\gamma-1} + \rho_-^\gamma}{\rho_1 - \rho_-}. \tag{3.21}$$

Define $x = \frac{\rho_1}{\rho_-} > 1$ and introduce $f(x) = (\gamma-1)x^\gamma - \gamma x^{\gamma-1} + 1$; then one can easily see that

$$\sigma_1 - \omega_2 = \frac{\rho_-^\gamma}{\rho_1 - \rho_-} f(x).$$

Obviously, we have $f'(x) = (\gamma-1)\gamma x^{\gamma-2}(x-1) > 0$ for $x > 1$, which gives $f(x) > f(1) = 0$ and then $\sigma_1 > \omega_2$. Thus S_1 will overtake R_2 in finite time and the intersection (x_2, t_2) can be calculated by

$$\begin{cases} x_2 + \varepsilon = \Big(u_- - \dfrac{\rho_1(\rho_1^{\gamma-1} - \rho_-^{\gamma-1})}{\rho_1 - \rho_-}\Big)t_2, \\ x_2 - x_1 = (u_1 - (\gamma-1)\rho_1^{\gamma-1})(t_2 - t_1), \end{cases} \tag{3.22}$$

in which (x_1, t_1) has the same representation as (3.12).

Hence, (x_2,t_2) can be expressed as

$$(x_2,t_2)=\left(\frac{2\varepsilon[\rho_1^{\gamma-1}(\rho_1-\rho_-)-\rho_1^{\gamma}(\rho_1^{\gamma-1}-\rho_-^{\gamma-1})]}{\rho_m^{\gamma-1}[(\gamma-1)\rho_1^{\gamma}-\gamma\rho_-\rho_1^{\gamma-1}+\rho_-^{\gamma}]}-\varepsilon,\frac{2\varepsilon\rho_1^{\gamma-1}(\rho_1-\rho_-)}{\rho_m^{\gamma-1}[(\gamma-1)\rho_1^{\gamma}-\gamma\rho_-\rho_1^{\gamma-1}+\rho_-^{\gamma}]}\right).$$

When $t>t_2$, S_1 begins to penetrate R_2, and the shock wave $x=x(t)$ during the process of penetration satisfies

$$\begin{cases}\dfrac{dx}{dt}=u_--\dfrac{\rho(\rho^{\gamma-1}-\rho_-^{\gamma-1})}{\rho-\rho_-},\\ x-\hat{x}=(u-(\gamma-1)\rho^{\gamma-1})(t-\hat{t}),\\ u-u_1=\rho_1^{\gamma-1}-\rho^{\gamma-1},\\ x(t_2)=x_2,\quad 0\le\rho\le\rho_1,\end{cases}\tag{3.23}$$

in which $(\hat{x},\hat{t})$ are the translation points from R_1 to R_2 and can be calculated by (3.13), but here $\rho_3\le\rho\le\rho_m$.

Similarly, by differentiating (3.23) with respect to t along $x=x(t)$ and noting that $u-u_-=-\rho^{\gamma-1}+\rho_-^{\gamma-1}$, we finally obtain, for $\rho>\rho_-$,

$$\frac{d\rho}{dt}=-\frac{(\gamma-1)\rho^{\gamma}-\gamma\rho_-\rho^{\gamma-1}+\rho_-^{\gamma}}{(\gamma-1)\gamma\rho^{\gamma-2}(\rho-\rho_-)(t-\hat{t})}<0,\tag{3.24}$$

$$\begin{aligned}\frac{d^2x}{dt^2}&=-\frac{(\gamma-1)\rho^{\gamma}-\gamma\rho_-\rho^{\gamma-1}+\rho_-^{\gamma}}{(\rho-\rho_-)^2}\cdot\frac{d\rho}{dt}\\&=\frac{((\gamma-1)\rho^{\gamma}-\gamma\rho_-\rho^{\gamma-1}+\rho_-^{\gamma})^2}{(\gamma-1)\gamma\rho^{\gamma-2}(\rho-\rho_-)^3(t-\hat{t})}>0,\end{aligned}\tag{3.25}$$

which means that the shock wave accelerates during the process of penetration.

Integrating (3.24) yields

$$\ln\frac{t-\hat{t}}{t_2-\hat{t}}=-\int_{\rho_1}^{\rho}\frac{(\gamma-1)\gamma\rho^{\gamma-2}(\rho-\rho_-)}{(\gamma-1)\rho^{\gamma}-\gamma\rho_-\rho^{\gamma-1}+\rho_-^{\gamma}}d\rho,\tag{3.26}$$

and we see that $t\to\infty$ as $\rho\to\rho_-$. Therefore, S_1 cannot penetrate over R_2 forever and the propagating speed of the shock wave will tend to $u_--(\gamma-1)\rho_-^{\gamma-1}$ as $t\to\infty$.

In brief, when $t\to\infty$, the solution can be expressed as

$$(u_-,\rho_-)+R+Vac+J_2+(u_+,\rho_+).$$

Remark 1. The curve passing through (u_-,ρ_-) is below the one passing through (u_m,ρ_m) if $u_-+\rho_-^{\gamma-1}<u_m+\rho_m^{\gamma-1}$ (i.e., $u_*<u_\sim$) in the (u,ρ) plane; otherwise the situation is opposite. Here we select the situation $u_*<u_\sim$ to discuss, and the other can be dealt with similarly.

4. Stability analysis and comparison. In this section, let us first consider whether the limits of the perturbed solutions of (1.2) and (1.3) are the corresponding Riemann solutions of (1.2) and (2.1). It is obviously true when the vacuum is not involved. On the other hand, if the vacuum is involved, let us take Case 3.1 as an example to study the limit situations of the above perturbed solutions for details.

In Case 3.1, it can be easily derived from (3.2) that $(x_1, t_1) \to (0, 0)$ as $\varepsilon \to 0$; thus the three points $(-\varepsilon, 0)$, $(\varepsilon, 0)$, and (x_1, t_1) coincide with each other in the limit situation. If $u_+ \geq u_* > u_-$ (see Figure 3.1), as $\varepsilon \to 0$, the intermediate state (u_2, ρ_2) disappears while SJ and J_2 unify into one contact discontinuity J since SJ and J_2 propagate with the same speed u_+. So the structure of the solution tends to $(u_-, \rho_-) + R + Vac + J + (u_+, \rho_+)$ as $\varepsilon \to 0$. Otherwise, if $u_+ < u_*$ (see Figure 3.2), we can also see that $(x_2, t_2) \to (0, 0)$ as $\varepsilon \to 0$ from (3.3) and the vacuum state disappears in the limit situation. Furthermore, S_2 cannot penetrate over R_1 for $u_- \leq u_+$, and the limit of the perturbed solution is $(u_-, \rho_-) + R + (u_3, \rho_3) + J + (u_+, \rho_+)$. Otherwise, S_2 is able to penetrate the whole of R_1 for $u_+ < u_-$, and the limit situation is $(u_-, \rho_-) + S + (u_3, \rho_3) + J + (u_+, \rho_+)$.

Therefore, in Case 3.1 the Riemann solutions are claimed to be stable for such perturbations with the Riemann data (2.1). The other cases can be analyzed similarly. Hence, we can see that the limits of the solutions of the perturbed Riemann problems (1.2) and (1.3) are exactly the solutions of the corresponding Riemann problems (1.2) and (2.1), which shows the stability of the Riemann solutions with respect to the small perturbations of the initial data in this particular situation.

In [2], Aw and Rascle considered the Riemann problem when the left (or right) state is the vacuum state and they fixed the right (or left) state and slightly perturbed the left (or right) state so that $0 < \rho_- \ll 1$ (or $0 < \rho_+ \ll 1$). They discovered that a big oscillation appeared and the solution dramatically changed under their small perturbations in some cases. Thus, the Riemann solution displays the instability when one of the Riemann data is near the vacuum under their perturbations. In [11], Lebacque, Mammer, and Haj-Salem extended properly the fundamental diagram (equilibrium speed-density relationship) in a suitable fashion to solve the Riemann problem for all initial conditions and to avoid irregular behavior at the low densities pointed out by Aw and Rascle.

If we adopt the perturbation such as (1.3), the stability of the Riemann solution can be obtained under the assumption $\rho_\pm > 0$ in section 3. Indeed, this stability can also be arrived at when one of the Riemann data is the vacuum state. Now, in order to compare with the results in [2], let us reconsider two interesting and typical examples: the perturbation (iv) of case 5 and the perturbation (vii) of case 4 (both in [2]).

Case 4.1. Let us first consider the perturbation (iv) of case 5 in [2]. In this case, the Riemann problem has no traffic on the left: $\rho_- = 0$, and the initial data on the right satisfies $u_+ < u_- < u_\star = u_+ + \rho_+^{\gamma-1}$. According to [2], for $\rho_- = 0$, the Riemann problem can be connected by a 2-contact discontinuity directly and the wave of the first family disappears. The shock with the large amplitude presents, and the Riemann solution is obviously unstable under the perturbation proposed by Aw and Rascle. But we can see that the Riemann solution is still stable if we take the perturbation (u_m, ρ_m) in the interval $(-\varepsilon, \varepsilon)$. According to the value of u_m, we divide our discussion into the following two subcases.

Subcase 4.1.1. If $u_+ < u_m$, then J_1 emerges from $(-\varepsilon, 0)$ and S_1 and J_2 start from $(\varepsilon, 0)$ (see Figure 4.1). Like in Case 3.1, J_1 must overtake S_1 in finite time and they unify into a contact discontinuity J_3 with the velocity u_+.

Subcase 4.1.2. If $u_+ > u_m$, then J_1 emerges from $(-\varepsilon, 0)$ and R_1 and J_2 emanate from $(\varepsilon, 0)$ (see Figure 4.2). Similarly to Case 3.2, J_1 will penetrate R_1 completely in finite time and then be denoted by J_3 with the velocity u_+, and R_1 will become the vacuum state at the same time.

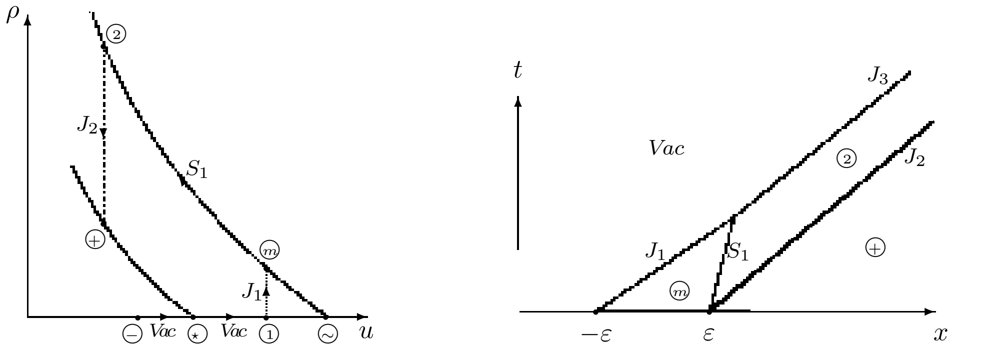


FIG. 4.1.

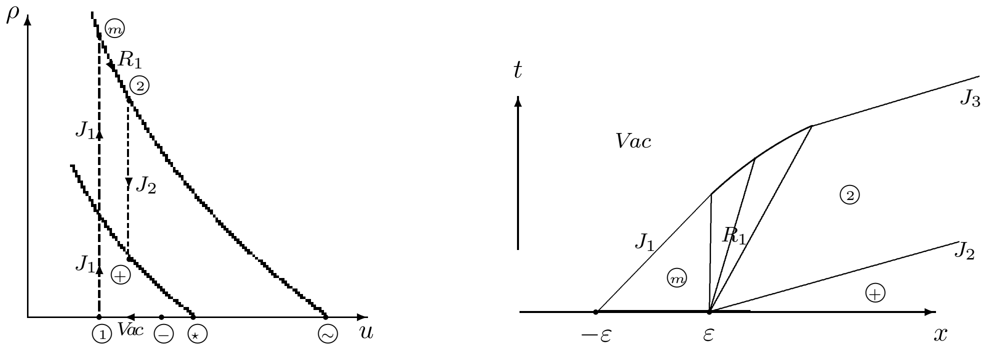


FIG. 4.2.

Remark 2. If $\rho_- = 0$ and $u_+ < u_-$, we can also believe that the Riemann solution consists of a 1-shock S connecting $(u_-, 0)$ and $(u_+, \sqrt[\gamma-1]{u_- - u_+})$, followed by a 2-contact discontinuity J connecting $(u_+, \sqrt[\gamma-1]{u_- - u_+})$ and (u_+, ρ_+). But S and J propagate at the same speed u_+; thus they coalesce into a new wave SJ and the intermediate state $(u_+, \sqrt[\gamma-1]{u_- - u_+})$ disappears in the (x, t) plane. The new wave SJ has the same properties as J and can also be regarded as J. Otherwise, if $\rho_- = 0$ and $u_+ > u_-$, it can also be believed that the Riemann solution consists of a fake vacuum wave connecting the two vacuum states $(u_-, 0)$ and $(u_+, 0)$ and then followed by a contact discontinuity J connecting $(u_+, 0)$ and (u_+, ρ_+).

Case 4.2. Let us now come back to the perturbation (vii) of case 4 in [2]. In this case, the Riemann data satisfies $\rho_+ = 0$ and $u_+ < u_-$. Based on [2], for $\rho_+ = 0$, the Riemann solution consists only of a rarefaction wave and there is no need to add a contact discontinuity. If we employ the perturbation in [2], the perturbed solution is still more dramatically different from the original one in that a (possible large) shock wave and a large contact discontinuity appear. However, we can see that the Riemann solution is still stable if we adopt the perturbation (1.3). Our discussion is also divided into the following two subcases according to the value of u_m.

Subcase 4.2.1. If $u_- < u_m$, then R_1 and J_1 generate from $(-\varepsilon, 0)$ and R_2 emanates from $(\varepsilon, 0)$. If $u_m \leq u_*$, as in Case 3.4, the vacuum state will form ahead of R_3 when one of the states in R_2 becomes (u_2, ρ_2). J_1 penetrates R_2 and has $x - \varepsilon = u_\sim t$ as its

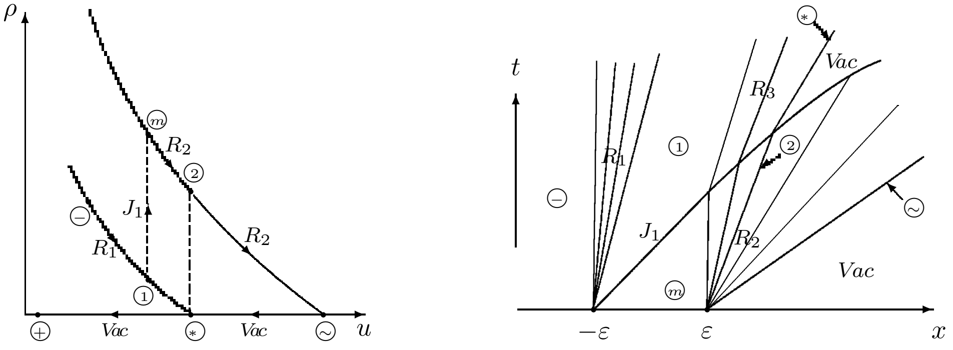


FIG. 4.3. $(u_+ <)u_- < u_m \leq u_*$.

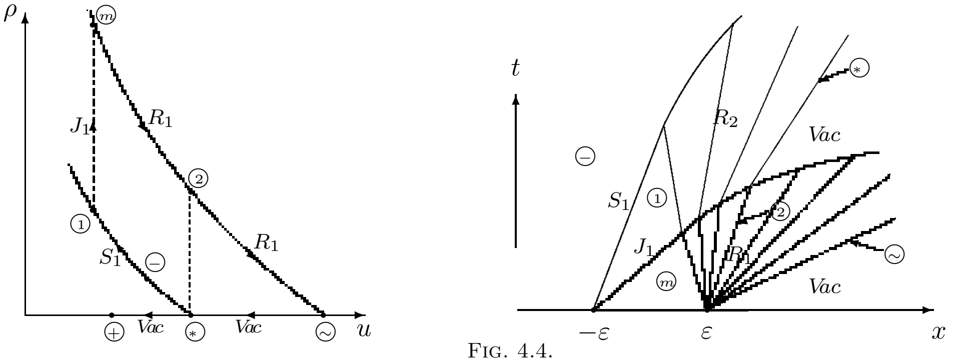


FIG. 4.4.

asymptote, which finally disappears at infinity for both sides are the vacuum states (see Figure 4.3). For the remaining case $u_m > u_*$, the conclusion is analogous and the difference lies in that R_3 disappears and the intermediate state between R_1 and J_1 becomes the vacuum state.

Subcase 4.2.2. If $u_- > u_m$, then S_1 and J_1 generate from $(-\varepsilon, 0)$ and R_1 emits from $(\varepsilon, 0)$. With the same reasoning as before, J_1 penetrates R_1 and has the wave front in R_1 as its asymptote; eventually it disappears in the vacuum as $t \to \infty$. For S_1, it cannot penetrate R_2 completely and its speed tends to $u_- - (\gamma - 1)\rho_-^{\gamma-1}$ in the end (see Figure 4.4).

Remark 3. If $\rho_+ = 0$ and $u_+ < u_-$ (or $u_- < u_+ < u_*$), we can also believe the Riemann solution consists of a 1-shock wave S (or a 1-rarefaction wave R) connecting (u_-, ρ_-) and $(u_+, \sqrt[\gamma-1]{u_- + \rho_-^{\gamma-1} - u_+})$ and then followed by a 2-contact discontinuity J connecting $(u_+, \sqrt[\gamma-1]{u_- + \rho_-^{\gamma-1} - u_+})$ and $(u_+, 0)$. Otherwise, if $\rho_+ = 0$ and $u_+ > u_*$, we can believe that the Riemann solution consists of a 1-rarefaction wave R connecting (u_-, ρ_-) and $(u_*, 0)$ and then followed by a fake vacuum wave connecting the two vacuum states $(u_*, 0)$ and $(u_+, 0)$.

By letting $\varepsilon \to 0$, it is easy to see that the Riemann solutions are stable under the perturbations (1.3) in the above two cases. The other cases in [2] can be treated in the same way, and the results are also identical with our assertions. It should be pointed out that the perturbations adopted in this paper are local perturbations, which are obviously different from the perturbations proposed by Aw and Rascle. This may be used to explain the following traffic situation: the perturbation of traffic status in a small range will come back soon.

5. Conclusion. So far, the discussion for all kinds of interactions has been completed. We have globally constructed the solutions for the perturbed initial value problem (1.2) and (1.3). From the above analysis, we can find that the asymptotic behavior of the perturbed solutions is governed completely by the states $(u_\pm, \rho_\pm)$. That is, the elementary waves in the time-asymptotic solutions consist of R and J for $u_+ > u_-$, or S and J for $u_+ < u_-$. Particularly, for $u_- < u_* < u_+$, the vacuum states are involved as the intermediate states between R and J in the time-asymptotic solutions. Thus we can draw the conclusion that the Riemann solutions of (1.2) and (2.1) are stable with respect to the small perturbations of the initial data in this particular situation.

REFERENCES

[1] A. Aw, A. Klar, A. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader model*, SIAM J. Appl. Math., 63 (2002), pp. 259–278.

[2] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.

[3] F. Berthelin, P. Degond, M. Delitata, and M. Rascle, *A model for the formation and evolution of traffic jams*, Arch. Ration. Mech. Anal., 187 (2008), pp. 185–220.

[4] T. Chang and L. Hsiao, *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Pitman Monographs and Surveys in Pure and Applied Mathematics 41, Longman Scientific and Technical, Harlow, UK, 1989.

[5] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss., Springer-Verlag, Berlin, 2000.

[6] C. Daganzo, *Requiem for second order fluid approximations of traffic flow*, Transportation Res. B, 29 (1995), pp. 277–286.

[7] M. Garavello and B. Piccoli, *Traffic flow on a road network using the Aw–Rascle model*, Comm. Partial Differential Equations, 31 (2006), pp. 243–275.

[8] J. M. Greenberg, *Extensions and amplifications of a traffic model of Aw and Rascle*, SIAM J. Appl. Math., 62 (2001), pp. 729–745.

[9] J. M. Greenberg, A. Klar, and M. Rascle, *Congestion on multilane highways*, SIAM J. Appl. Math., 63 (2003), pp. 818–833.

[10] M. Herty and M. Rascle, *Coupling conditions for a class of second-order models for traffic flow*, SIAM J. Math. Anal, 38 (2006), pp. 595–616.

[11] J. P. Lebacque, S. Mammer, and H. Haj-Salem, *The Aw–Rascle and Zhang's model: Vacuum problems, existence and regularity of the solutions of the Riemann problem*, Transportation Res. B, 41 (2007), pp. 710–721.

[12] T. P. Liu and J. Smoller, *On the vacuum state for isentropic gas dynamic equations*, Adv. in Appl. Math., 1 (1980), pp. 345–359.

[13] S. Moutari and M. Rascle, *A hybrid Lagrangian model based on the Aw–Rascle traffic flow model*, SIAM J. Appl. Math., 68 (2007), pp. 413–436.

[14] D. Serre, *Systems of Conservation Laws,* 1 *and* 2, Cambridge University Press, Cambridge, UK, 1999 and 2000.

[15] C. Shen and M. Sun, *Formation of Delta-Shocks and Vacuum States in the Vanishing Pressure Limit of Solutions to the Aw–Rascle Model*, submitted.

[16] W. Sheng, M. Sun, and T. Zhang, *The generalized Riemann problem for a scalar nonconvex Chapman–Jouguet combustion model*, SIAM J. Appl. Math., 68 (2007), pp. 544–561.

[17] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.

[18] M. Sun and W. Sheng, *The ignition problem for a scalar nonconvex combustion model*, J. Differential Equations, 231 (2006), pp. 673–692.

[19] B. Temple, *Systems of conservation laws with coinciding shock and rarefaction curves*, Contemp. Math., 17 (1983), pp. 143–151.

[20] M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transportation Res. B, 36 (2002), pp. 275–290.

SIAM J. Appl. Math.
Vol. 69, No. 6, pp. 1559–1579

DRIFT-DIFFUSION PAST A CIRCLE: SHADOW REGION ASYMPTOTICS[*]

SEAN LYNCH[†] AND CHARLES KNESSL[†]

Abstract. We consider the steady state concentration of some diffusing substance, subject to a uniform drift field, past a circular obstacle. We obtain some exact representations of the concentration profile of the substance exterior to the obstacle. These representations are particularly useful for studying the solution in ranges of space where the concentration is very small (the "shadow" regions). We assume then that the drift dominates diffusion and obtain various asymptotic expansions in the shadow regions.

Key words. convection, diffusion, shadow, asymptotics

AMS subject classifications. 41A60, 35J25, 33C10

DOI. 10.1137/080717109

1. Introduction.

A very basic problem in classical mathematical physics is the study of a diffusing substance drifting past an obstacle. Such problems occur in many applications, including chromatography, groundwater flow, electrophoresis, and sedimentation. We will briefly discuss several applications in the paragraphs below.

Perhaps the simplest geometry to consider for such problems is that of a circle in $\mathbf{R}^2$ (or circular cylinder in $\mathbf{R}^3$). The problem of steady state linear drift-diffusion past a circle was analyzed exactly by Philip, Knight, and Waechter in [1], where it was used to model unsaturated seepage of groundwater past a circular cylindrical impermeable obstacle such as a rock, tunnel, or for their purposes a subterranean hole. The air contained in a subterranean hole is at a greater pressure than that of seeping water originating from the soil's surface (see [5]). Thus water may enter subterranean holes only under circumstances that allow water pressure on the surface of the hole to sufficiently increase. Factors contributing to these circumstances include the seepage velocity and the size and shape of the hole.

We note that the drift is caused by gravity in this application, which of course exists throughout the spatial region (including the inside of the circular obstacle). Capillarity provides the diffusive effects, and $\alpha L/2$, which we call c, measures the relative strength of the drift to the diffusion, where L is a characteristic length and α is the sorptive number from hydrogeology (which measures the capacity of a porous medium for capillary uptake of a fluid). The concentration of groundwater, which we call p, is the quantity of interest.

Similar mathematical models were used in connection with the Brazil nut effect [11], [12], [13]. Here a container of different sized particles subjected to shaking becomes segregated, with larger particles rising to the top of the container. We can view the larger nut as the obstacle, the smaller nuts as the diffusing substance, gravity as the drift, and shaking as diffusion. In [13] qualitative comparisons were drawn between the Brazil nut problem and the groundwater seepage problem in [1].

[*] Received by the editors February 28, 2008; accepted for publication (in revised form) October 31, 2008; published electronically March 11, 2009. This work was partially supported by NSF grant DMS-05-03745 and NSA grant H 98230-08-1-0102.

http://www.siam.org/journals/siap/69-6/71710.html

[†] University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (slynch5@uic.edu, knessl@uic.edu).

We can also imagine other applications of this model. For example, we can consider a sea of charged particles subject to a uniform electromagnetic field in the presence of a fixed circular obstacle. The uniform electromagnetic field causes the drift, local particle interactions cause the diffusive effects, and the quantity of interest, p, now gives the concentration of particles. Another physical situation that could correspond to our model is the concentration of Brownian particles in still air, such as dust particles, subject to gravity and with the collection of particles modeled as a continuum.

Considering the original application, the solution obtained in [1] gave the steady state concentration of groundwater as a Fourier series, with the Fourier coefficients characterized as infinite series involving the modified Bessel functions $I_\nu(\cdot)$ and $K_\nu(\cdot)$. Due to the complexity of the solution and the resulting numerical difficulties which are exacerbated when c is large, an asymptotic solution for $c \gg 1$ is desirable, and this is the topic of this paper. The asymptotic solution will also provide qualitative information about the concentration, such as the manner in which it transitions from being constant in the far field to being small directly beneath the obstacle.

In [2] we obtained another form of the solution given in [1] and studied various asymptotic properties. The asymptotic analysis assumes that c is large. We assume that (i) the drift field is of uniform strength, (ii) the field exists in all of $\mathbf{R}^2$ (including inside the circular obstacle), (iii) there is a uniform concentration of particles infinitely far from the obstacle, and (iv) the normal component of the flux of particles vanishes on the obstacle's exterior boundary. With these assumptions we can divide the exterior of the obstacle into an "illuminated region" ($\{r = \sqrt{x^2+y^2} > 1,\ |y| > 1\} \cup \{r > 1,\ -1 \le y \le 1,\ x < 0\}$) and a complementary "shadow region" $\Omega_s = \{(x,y) \mid r > 1,\ -1 < y < 1,\ x > 0\}$. We note that this definition excludes the shadow side of the obstacle, where $r \approx 1$ and $x > 0$. Here we scale the spatial variables so that the obstacle is inside the unit circle $r = 1$, and the uniform concentration at $r = \infty$ is taken as unity. For large c the concentration inside the illuminated region will be asymptotically equal to one, while that in the shadow region will be exponentially small. These definitions of shadow and illuminated regions are useful qualitatively even for moderate c.

In [2] we gave asymptotic results for the illuminated region and also where the shadow boundaries (defined by $y = \pm 1,\ x > 0$) meet the obstacle. This corresponds to $(x, y) = (0, \pm 1)$, and in this range the particle concentration is asymptotically large, of the order $O(c^{2/3})$. In this paper we shall obtain detailed asymptotic results in the shadow region, including the shadow boundaries and the shadow side of the obstacle (where $r \approx 1$ and $0 < x < 1$).

In particular we shall obtain the exponentially small concentration profile in the shadow region, Ω_s. Then we will construct "boundary layer" expansions near the obstacle's shadow side, where $r - 1 = O(c^{-1})$ and $r - 1 = O(c^{-2/3})$. Thus, two nested layers are needed. In the shadow boundary we will consider the two scales $y - 1 = O(c^{-1/3})$ and $y - 1 = O(c^{-1/2})$. On the latter scale the concentration of particles is large, of the order $O(\sqrt{c})$, and follows a Gaussian profile in the similarity variable $\sqrt{c}(y-1)/\sqrt{x}$.

In studying the shadow region asymptotically it proves necessary to have a representation that is different from that of the Fourier series in [1] and [2], and we shall obtain such a representation. It does not seem possible to obtain the shadow region asymptotics from the Fourier series.

There has been much previous work on related problems, and we do not make any attempt to survey it here. Some related papers are those of Chapman, Lawry, and

Ockendon [3] and Cherepanov [4]. In [3] the authors analyzed temperature exterior to a slit, with a convecting fluid velocity field that avoids the slit. Thus in this case the field is not a constant drift field as it exists only exterior to the obstacle. The authors also use conformal mapping to treat more complicated obstacle geometries. In [4] the author obtained exact solutions for the case of a slit, in terms of Mathieu and modified Mathieu functions. Other exact solutions for simple geometries were obtained by Knight and Philip [6] for spherical obstacles and by Philip [7], [8], [9], [10].

This work assumes that the drift field is constant and exists everywhere in space. Thus the mathematical model would be appropriate when the drift is due to gravity or electromagnetic fields, but not for, say, a diffusing substance in a liquid (such as water) flowing past an obstacle. For the latter the convection field would be some given potential or viscous flow that would itself satisfy a boundary condition on the surface of the obstacle.

We mention that the asymptotic structure of drift-diffusion problems for large c has some similarities to scattering problems in the high frequency limit. However, there are important differences. For example, the scattering problem for a plane wave hitting a circle in $\mathbf{R}^2$ would have the shadow boundary correspond to the disappearance of the incoming plane wave via a Fresnel integral. But the drift-diffusion problems in this spatial range have to leading order a Gaussian behavior. A good survey of exact and asymptotic results for scattering by simple geometries can be found in [14].

The remainder of the paper is organized as follows. In section 2 we summarize our results, both exact and asymptotic. The exact results are derived in section 3 and the asymptotic ones in sections 4–6. Concluding remarks are given in section 7.

2. Summary of results.

We let $p(x, y)$ be the concentration of some substance that undergoes diffusion exterior to the unit circle $r = \sqrt{x^2 + y^2} = 1$ and is subject to a constant drift field in the $+x$ direction. Then the steady state concentration, p, satisfies the following linear PDE:

$$p_{xx} + p_{yy} - 2cp_x = \Delta p - 2cp_x = 0, \ r > 1. \tag{1}$$

Here $c > 0$ is a parameter measuring the ratio of drift to diffusion. We assume that the substance cannot penetrate the obstacle, which leads to a boundary condition at $r = 1$. The flux vector of the substance is $\mathbf{J} = (2cp - p_x, -p_y)$. The component of $\mathbf{J}$ normal to the obstacle must vanish, which leads to

$$(\cos\theta)p_x + (\sin\theta)p_y - 2c(\cos\theta)p = p_r - 2c\cos\theta p = 0, \ r = 1. \tag{2}$$

The other boundary condition we impose is far away from the obstacle. We assume that the concentration approaches a constant value, which we take to be one; hence

$$p(x, y) \to 1 \ , \ r = \sqrt{x^2 + y^2} \to \infty. \tag{3}$$

For convenience we define $q(x, y)$ via

$$p(x, y) = 1 + e^{cx} q(x, y). \tag{4}$$

Upon changing variables in (1)–(3) we see that q satisfies the Helmholtz equation

$$q_{xx} + q_{yy} = c^2 q, \ r > 1, \tag{5}$$

with the boundary conditions

$$(\cos\theta)q_x + (\sin\theta)q_y - c(\cos\theta)q = 2c(\cos\theta)e^{-cx},\ r = 1, \tag{6}$$

and

$$q(x,y) \to 0\ ,\ r = \sqrt{x^2+y^2} \to \infty. \tag{7}$$

In polar coordinates (6) becomes

$$q_r - c(\cos\theta)q = 2c(\cos\theta)e^{-c(\cos\theta)},\ r = 1. \tag{8}$$

The problem (1)–(3) was explicitly solved by Philip, Knight, and Waechter [1], and a different representation of the solution is given in [2], where it was shown that q in (4) is given by

$$q(r,\theta) = -\int_r^\infty e^{c(\cos\theta)(r-s)}\tilde{q}(s,\theta)ds, \tag{9}$$

where

$$\tilde{q}(s,\theta) = 2c(\cos\theta)Q_1(s,\theta) - Q_2(s,\theta) \tag{10}$$

with

$$\begin{aligned} Q_1(s,\theta) &= \sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{K_\nu(cs)}{K_\nu(c)}I_{|\nu|}(c)e^{i\nu(\theta+(2m+1)\pi)}d\nu \\ &= 2\mathrm{Re}\left[\sum_{m=-\infty}^{\infty}\int_0^\infty \frac{K_\nu(cs)}{K_\nu(c)}I_\nu(c)e^{i\nu(\theta+(2m+1)\pi)}d\nu\right] \end{aligned} \tag{11}$$

and

$$Q_2(s,\theta) = \sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\nu}{K_\nu(c)}e^{i\nu(\theta+(2m+1)\pi)}\left[\frac{K_{\nu+1}(cs)}{K_{\nu+1}(c)} - \frac{K_{\nu-1}(cs)}{K_{\nu-1}(c)}\right]d\nu. \tag{12}$$

Here K and I are modified Bessel functions.

We give below an alternate form for p, which will prove useful for calculations in the "shadow region," $\Omega_s = \{(x,y)\mid r>1,\ -1<y<1,\ x>0\}$.

THEOREM 2.1. *The solution $p(x,y)$ to (1)–(3) is given by*

$$\begin{aligned} p = \mathrm{Im}\Bigg\{\int_r^\infty e^{c(\cos\theta)(2r-s)}\sum_{k=0}^{\infty}\frac{-4\pi\mathrm{csch}(\omega_k\pi)}{\dot{K}_{i\omega_k}(c)K_{i\omega_k+1}(c)}\Big[i\omega_k\cosh(\omega_k\theta)K_{i\omega_k+1}(cs) \\ + \left(\omega_k\sin\theta\sinh(\omega_k\theta) + i\omega_k\cos\theta\cosh(\omega_k\theta) - i\sin\theta\sinh(\omega_k\theta)\right)K_{i\omega_k}(cs)\Big]\Bigg\}ds. \end{aligned} \tag{13}$$

Here $\dot{K}$ is used to denote the derivative of K with respect to order, and the ω_k are real positive solutions to $K_{i\omega_k}(c) = 0$, ordered as $0<\omega_0<\omega_1<\cdots$.

Another form of the solution is given by

$$\begin{aligned} p = \mathrm{Im}\Bigg\{\int_{-\infty}^{\infty}\sum_{k=0}^{\infty}\frac{-2\pi c^{-1}\exp\left[i\omega_k t\right]\exp\left[cr\left(\cos\theta - \cosh t\right)\right]}{(\cos\theta+\cosh t)\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)K_{i\omega_k+1}(c)} \\ \times\left[i\omega_k\cosh(\omega_k\theta)e^t + i\omega_k\cos\theta\cosh(\omega_k\theta) + (\omega_k - i)\sin\theta\sinh(\omega_k\theta)\right]dt\Bigg\}. \end{aligned} \tag{14}$$

While (13) and (14) are exact forms of the solution, their usefulness is still numerically constrained to situations involving moderate values of c, as are the forms given in [1] and by (4), (9)–(12). Also, their complexity is such that they do not yield qualitative insights into the solution.

We next give various asymptotic approximations to $p(x, y)$ for $c \to \infty$. The results apply in the shadow region with the following scaling:

(A)	Shadow region, Ω_s^+:	$r > 1,\ x > 0,\ 0 < y < 1.$
(B)	Intermediate shadow boundary:	$y - 1 = O(c^{-1/3}),$
	Inner shadow boundary:	$y - 1 = O(c^{-1/2}).$
(C)	Intermediate layer near obstacle:	$r - 1 = O(c^{-2/3}),\ 0 < y < 1,$
	Inner layer near obstacle:	$r - 1 = O(c^{-1}),\ 0 < y < 1.$

We summarize our results in Theorems 2.2–2.4 below. In view of the symmetry $p(x, y) = p(x, -y)$ it suffices to consider $y \geq 0$.

THEOREM 2.2. *In the domain Ω_s^+, we have*

$$\begin{aligned} p = {} & c^{5/6}\, \frac{2^{1/6} T(r,\theta)}{\sqrt{\pi}\left[\mathrm{Ai}'(r_0)\right]^2}\ \exp\left[c\left(r\cos\theta + \theta - \sin^{-1}\left(\frac{1}{r}\right) - \sqrt{r^2-1}\right)\right] \\ & \times\ \left(r^2-1\right)^{-1/4} \exp\left[c^{1/3} 2^{-1/3} |r_0| \left(\theta - \sin^{-1}\left(\frac{1}{r}\right)\right)\right] \\ & \times\ \left\{1 + c^{-1/3}\left[\gamma_0\left(\theta - \sin^{-1}\left(\frac{1}{r}\right)\right) - \frac{15\gamma_0}{\sqrt{r^2-1}}\right] + O(c^{-2/3})\right\}, \end{aligned} \tag{15}$$

where

$$T(r,\theta) = \frac{1 - r\sin\theta}{r\cos\theta + \sqrt{r^2-1}}, \qquad \gamma_0 = \frac{2^{-2/3} r_0^2}{30}, \tag{16}$$

and $r_0 = \max\{z : \mathrm{Ai}(z) = 0\} = -2.33811\ldots$.

If $-1 < y < 0$, we should replace θ by $-\theta$. If y is small, specifically $y = O(c^{-1})$ (thus $\theta = O(c^{-1})$), we must *add* the results with θ and $-\theta$.

THEOREM 2.3. *For the shadow boundary intermediate layer, where $y - 1 = O(c^{-1/3})$ and $x > 0$, we set $y - 1 = c^{-1/3} Y$ with $Y = O(1)$. For $Y < 0$ we obtain*

$$\begin{aligned} p = {} & c^{1/2} \frac{2^{-5/6}}{\sqrt{\pi}} \exp\left[-c^{1/3}\frac{Y^2}{2x}\right] \exp\left[\frac{Y^3}{6x^3}\right] \sum_{k=0}^{\infty} \frac{e^{A_k Y/x}}{\left[\mathrm{Ai}'(r_k)\right]^2} \\ & \times \left\{R_0(x,Y,k) + c^{-1/3} R_1(x,Y,k) + c^{-2/3} R_2(x,Y,k) + O(c^{-1})\right\}, \end{aligned} \tag{17}$$

where

$$R_0 = -\frac{Y}{x^{3/2}}, \tag{18}$$

$$R_1 = \left(\frac{Y}{x}\right)^5 \frac{1-x^2}{8x^{3/2}} + \left(\frac{Y}{x}\right)^3 \frac{A_k}{2x^{3/2}} + \left(\frac{Y}{x}\right)^2 \frac{1}{x^{3/2}} + \left(\frac{Y}{x}\right) \frac{A_k^2}{2x^{3/2}} + \frac{A_k}{x^{3/2}}, \tag{19}$$

$$\begin{aligned} R_2 = {} & \left(\frac{Y}{x}\right)^9 \frac{-(x^2-1)^2}{128x^{5/2}} + \left(\frac{Y}{x}\right)^7 \frac{A_k(x^2-1)}{16x^{5/2}} + \left(\frac{Y}{x}\right)^6 \frac{6x^2-5}{20x^{5/2}} \\ & + \left(\frac{Y}{x}\right)^5 \frac{A_k^2(x^2-3)}{16x^{5/2}} + \left(\frac{Y}{x}\right)^4 \frac{A_k(11x^2-27)}{24x^{5/2}} + \left(\frac{Y}{x}\right)^3 \frac{4x^2-11-2A_k^2}{8x^{5/2}} \\ & + \left(\frac{Y}{x}\right)^2 \frac{-A_k^2(x^2+45)}{30x^{5/2}} + \left(\frac{Y}{x}\right)\left(\frac{-A_k^4}{8x^{5/2}} - \frac{A_k(16x^2+30)}{15x^{5/2}}\right) - \frac{A_k^3+1}{2x^{5/2}}, \end{aligned} \tag{20}$$

and $A_k = 2^{-1/3}|r_k|$. *The* r_k *are the real negative zeros of the Airy function (thus* $\mathrm{Ai}(r_k) = 0$*) such that* $|r_0| < |r_1| < |r_2| < \cdots$. *We can also write* (17)–(20) *in an alternate integral form that will apply also for* $Y > 0$. *We have*

$$p = c^{1/2}2^{-1/6}\pi^{-1/2}\exp\left[-c^{1/3}\frac{Y^2}{2x}\right]\exp\left[\frac{Y^3}{6x^3}\right] \\ \times\left\{\mathcal{F}_0 + c^{-1/3}\mathcal{F}_1 + c^{-2/3}\mathcal{F}_2 + O(c^{-1})\right\}, \tag{21}$$

where

$$\mathcal{F}_0 = \frac{x^{-1/2}}{2\pi i}\int_C e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)}, \tag{22}$$

$$\begin{aligned}\mathcal{F}_1 = {} & \frac{x^{1/2}}{2\pi i}\int_C\left(\frac{Y^4}{8x^4}\right)e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)} \\ & - \frac{x^{-3/2}}{2\pi i}\int_C\left(\frac{Y^4}{8x^4} + \frac{Y}{2x} - z\frac{Y^2}{2x^2} + z^2\frac{1}{2}\right)e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)},\end{aligned} \tag{23}$$

$$\begin{aligned}\mathcal{F}_2 = {} & \frac{1}{2^{2/3}x^{1/2}}\sum_{k=0}^{\infty}\frac{e^{A_kY/x}}{\left[\mathrm{Ai}'(r_k)\right]^2} + \frac{x^{3/2}}{2\pi i}\int_C\left(\frac{Y^8}{128x^8}\right)e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)} \\ & + \frac{x^{-1/2}}{2\pi i}\int_C\left[\frac{-Y^8}{64x^8} - \frac{19Y^5}{80x^5} - \frac{Y^2}{6x^2} + z\left(\frac{Y^6}{16x^6} + \frac{Y^3}{3x^3} - 1\right)\right. \\ & \qquad \left. + z^2\left(\frac{-Y^4}{16x^4} + \frac{Y}{30x}\right)\right]e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)} \\ & + \frac{x^{-5/2}}{2\pi i}\int_C\left[\frac{Y^8}{128x^8} + \frac{3Y^5}{16x^5} + \frac{5Y^2}{8x^2} + z\left(\frac{-Y^6}{16x^6} - \frac{3Y^3}{4x^3} - \frac{1}{2}\right)\right. \\ & \qquad \left. + z^2\left(\frac{3Y^4}{16x^4} + \frac{3Y}{4x}\right) - z^3\frac{Y^2}{4x^2} + z^4\frac{1}{8}\right]e^{-zY/x}\frac{dz}{\mathrm{Ai}^2(2^{1/3}z)}.\end{aligned} \tag{24}$$

The integration contour C *is the imaginary axis (*z *goes from* $-i\infty$ *to* $i\infty$*). Expressions* (21)–(24) *give a three-term approximation to* p *for* $Y < 0$ *and a two-term approximation to* $p - 1$ *for* $Y > 0$ *(here we use only* (21)–(23)*). However, for* $Y \approx 0$ *a different expansion is needed, which is given below.*

For the shadow boundary inner layer, where $y - 1 = O(c^{-1/2})$, *and* $x > 0$, *we set* $\Delta = c^{1/2}\,(y - 1)$ *and obtain*

$$p = c^{1/2}p^{(0)} + c^{1/3}p^{(1/3)} + c^{1/6}p^{(2/3)} + p^{(1)} + O(c^{-1/6}), \tag{25}$$

where

$$p^{(0)} = \frac{1}{\sqrt{2\pi x}}e^{-\Delta^2/(2x)}, \tag{26}$$

$$p^{(1/3)} = \left(\frac{\Delta}{x}\right)\frac{-2^{-1/3}C_1}{\sqrt{2\pi x}}e^{-\Delta^2/(2x)}, \tag{27}$$

$$p^{(2/3)} = \left(\frac{\Delta^2}{2x^2} - \frac{1}{2x}\right) \frac{2^{-2/3}C_2}{\sqrt{2\pi x}} e^{-\Delta^2/(2x)}, \tag{28}$$

$$\begin{aligned} p^{(1)} = {} & \left(\frac{\Delta^3}{6x^3} - \frac{\Delta}{2x^2}\right) \frac{1 - 2^{-1}C_3}{\sqrt{2\pi x}} e^{-\Delta^2/(2x)} \\ & + 1 - 2^{-1/2}\pi^{-1/2} e^{-\Delta^2/(2x)} \int_0^\infty \exp\left[-\frac{z\Delta}{\sqrt{x}} - \frac{z^2}{2}\right] dz. \end{aligned} \tag{29}$$

The constants C_j above are defined as

$$C_j = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{z^j}{[\mathrm{Ai}(z)]^2} dz, \tag{30}$$

and their approximate numerical values are $C_0 = 1$, $C_1 \approx -1.25512$, $C_2 \approx 1.06458$, $C_3 \approx 0.20315$.

We note that as $\Delta \to +\infty$ the expressions in (26)–(28) have Gaussian decay while $p^{(1)} \to 1$. Thus as we leave the inner shadow boundary we begin to see the uniform concentration at infinity as given in (3).

Next we give results that apply where the shadow region meets the obstacle's boundary. There are again two scales to consider.

Theorem 2.4. *For the intermediate layer near the obstacle, where $r - 1 = O(c^{-2/3})$, $0 < y < 1$, and $x > 0$, we set $\eta = c^{2/3}(r-1)$ and obtain*

$$\begin{aligned} p = c \; & \exp\Big[c\left(\cos\theta + \theta - \frac{\pi}{2}\right) + c^{1/3}2^{-1/3}|r_0|\left(\theta - \frac{\pi}{2}\right)\Big] \exp\left[c^{1/3}\eta\cos\theta\right] \\ & \times \left\{F(\theta)\mathrm{Ai}(\widehat{\eta}) + c^{-1/3}\Big[-2^{1/3}F'(\theta)\mathrm{Ai}'(\widehat{\eta}) + \gamma_0\left(\theta - \frac{\pi}{2}\right)F(\theta)\mathrm{Ai}(\widehat{\eta})\Big] + O(c^{-2/3})\right\}, \end{aligned} \tag{31}$$

where

$$F(\theta) = \frac{2\cos\theta}{\left[\mathrm{Ai}'(r_0)\right]^2 (1 + \sin\theta)} \tag{32}$$

and $\widehat{\eta} = 2^{1/3}(\eta - 2^{-1/3}|r_0|) = 2^{1/3}\eta + r_0$.

For the inner layer near the obstacle, where $r - 1 = O(c^{-1})$, $0 < y < 1$, and $x > 0$, we set $\xi = c(r-1)$ and obtain

$$\begin{aligned} p = {} & c^{2/3}\frac{2^{4/3}(1 + \xi\cos\theta)}{\mathrm{Ai}'(r_0)(1 + \sin\theta)} \exp\Big[c\left(\cos\theta + \theta - \frac{\pi}{2}\right) + c^{1/3}2^{-1/3}|r_0|\left(\theta - \frac{\pi}{2}\right)\Big] \\ & \times \exp\left[\xi\cos\theta\right]\left\{1 + c^{-1/3}\frac{2^{-2/3}}{30}r_0^2\left(\theta - \frac{\pi}{2}\right) + O(c^{-2/3})\right\}. \end{aligned} \tag{33}$$

If $-1 < y < 0$, we should replace θ by $-\theta$. If y is small, specifically $y = O(c^{-1})$ (thus $\theta = O(c^{-1})$), we must add the results with θ and $-\theta$.

3. The exact representation.

The main steps in deriving the alternate form of p given in Theorem 2.1 will be to write the integrals in (11) and (12) as residue series and explicitly evaluate the m-sums.

First consider (12). To transform the integral into a residue series we recall that the zeros of $K_\nu(c)$ all lie on the imaginary axis in the ν-plane, and that $K_\nu(c) = K_{-\nu}(c)$ for all $\nu \in \mathbf{C}$. Allowing ν to take on complex values, we see that

$$\mathrm{Re}\left[i\nu\left(\theta + (2m+1)\pi\right)\right] = -\mathrm{Im}(\nu)\left[\theta + (2m+1)\pi\right],$$

which is negative when $m < 0$ and $\mathrm{Im}(\nu) < 0$, or when $m \geq 0$ and $\mathrm{Im}(\nu) > 0$ since $|\theta| < \pi$. Thus for $m < 0$ we will close in the lower half-plane, where $\mathrm{Im}(\nu) < 0$, and for $m \geq 0$ we will close in the upper half-plane, where $\mathrm{Im}(\nu) > 0$. This yields

$$\begin{aligned} Q_2(s,\theta) = &-\sum_{m=-\infty}^{-1} \sum \mathrm{Res}\left\{ \frac{2\pi i\nu}{K_\nu(c)} e^{i\nu(\theta+(2m+1)\pi)} \left[\frac{K_{\nu+1}(cs)}{K_{\nu+1}(c)} - \frac{K_{\nu-1}(cs)}{K_{\nu-1}(c)} \right] \right\} \\ &+ \sum_{m=0}^{\infty} \sum \mathrm{Res}\left\{ \frac{2\pi i\nu}{K_\nu(c)} e^{i\nu(\theta+(2m+1)\pi)} \left[\frac{K_{\nu+1}(cs)}{K_{\nu+1}(c)} - \frac{K_{\nu-1}(cs)}{K_{\nu-1}(c)} \right] \right\}, \end{aligned} \tag{34}$$

where the two inner sums are over all singularities in the lower and upper half-planes, respectively. Here Res denotes residue. The singularities of (34) are the zeros of K_ν and $K_{\nu\pm1}$ which are located at $i\omega_k$ and $i\omega_k \pm 1$, $k = 0, 1, 2, \ldots$, in the upper half-plane and at $-i\omega_k$ and $-i\omega_k \pm 1$, $k = 0, 1, 2, \ldots$, in the lower half-plane. We will also simplify (34) by recognizing

$$\frac{K_{\nu+1}(cs)}{K_{\nu+1}(c)} - \frac{K_{\nu-1}(cs)}{K_{\nu-1}(c)} \tag{35}$$

as the difference between a function and its complex conjugate, for purely imaginary ν. Evaluating the residues leads to

$$\begin{aligned} Q_2(s,\theta) = &-2\pi i \sum_{m=1}^{\infty} \sum_{k=0}^{\infty} \left\{ \frac{2\omega_k}{\dot{K}_{i\omega_k}(c)} e^{\omega_k[\theta+(1-2m)\pi]} \mathrm{Im}\left[\frac{K_{i\omega_k+1}(cs)}{K_{i\omega_k+1}(c)} \right] \right. \\ &\left. + \frac{K_{i\omega_k}(cs)}{\dot{K}_{i\omega_k}(c)} e^{\omega_k[\theta+(1-2m)\pi]} 2\mathrm{Re}\left[e^{-i[\theta+(1-2m)\pi]} \frac{i\omega_k+1}{K_{i\omega_k+1}(c)} \right] \right\} \\ &+ 2\pi i \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} \left\{ \frac{-2\omega_k}{\dot{K}_{i\omega_k}(c)} e^{-\omega_k[\theta+(1+2m)\pi]} \mathrm{Im}\left[\frac{K_{i\omega_k+1}(cs)}{K_{i\omega_k+1}(c)} \right] \right. \\ &\left. + \frac{K_{i\omega_k}(cs)}{\dot{K}_{i\omega_k}(c)} e^{-\omega_k[\theta+(1+2m)\pi]} 2\mathrm{Re}\left[e^{i[\theta+(1+2m)\pi]} \frac{-i\omega_k-1}{K_{i\omega_k+1}(c)} \right] \right\}. \end{aligned} \tag{36}$$

The m-sums above are simply geometric series which we can evaluate, and after some rearrangement (36) becomes

$$\begin{aligned} Q_2(s,\theta) = 2\pi i \sum_{k=0}^{\infty} &\left\{ \frac{-2\omega_k \cosh(\omega_k\theta)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)} \mathrm{Im}\left(\frac{K_{i\omega_k+1}(cs)}{K_{i\omega_k+1}(c)} \right) \right. \\ &\left. + \frac{K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)} \mathrm{Re}\left(e^{\omega_k\theta} e^{-i\theta} \frac{1+i\omega_k}{K_{i\omega_k+1}(c)} + e^{-\omega_k\theta} e^{i\theta} \frac{1+i\omega_k}{K_{i\omega_k+1}(c)} \right) \right\}. \end{aligned} \tag{37}$$

We will now rewrite the expression inside $\mathrm{Re}(\cdot)$, replace $\mathrm{Re}(\cdot)$ by $\mathrm{Im}(i\cdot)$, and finally move the rest of the expression inside the Im which we can do since $(\dot{K}_{i\omega_k}(c))^{-1}$ and $2\pi i$ are purely imaginary; thus their product is real. We obtain

$$\begin{aligned} Q_2(s,\theta) = \mathrm{Im}&\left\{ \sum_{k=0}^{\infty} \frac{-4\pi \mathrm{csch}(\omega_k\pi)}{\dot{K}_{i\omega_k}(c) K_{i\omega_k+1}(c)} \left[i\omega_k \cosh(\omega_k\theta) K_{i\omega_k+1}(cs) \right. \right. \\ &\left. \left. + (1+i\omega_k)\left(\cos\theta\cosh(\omega_k\theta) - i\sin\theta\sinh(\omega_k\theta)\right) K_{i\omega_k}(cs) \right] \right\}. \end{aligned} \tag{38}$$

Now we turn our attention to Q_1 given by (11). Similarly to how we approached Q_2, for $m \geq 0$ we will rotate the contour by $+90°$, to a contour C_1, that goes from $\nu = 0$ to $\nu = i\infty$ with indentations to the right of all poles which lie at $\nu = i\omega_k$, $\omega_k > 0$. For $m \leq -1$ we rotate by $-90°$ to a contour C_2, that goes from $\nu = 0$ to $\nu = -i\infty$, again indented to the right of the poles which lie at $\nu = -i\omega_k$, $\omega_k > 0$. We thus obtain

$$\begin{aligned} Q_1(s,\theta) = 2\mathrm{Re}\Bigg\{ &\sum_{m=0}^{\infty} \int_{C_1} \frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \exp\Big[i\nu\left(\theta + (2m+1)\pi\right)\Big]\, d\nu \\ &+ \sum_{m=1}^{\infty} \int_{C_2} \frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \exp\Big[i\nu\left(\theta + (1-2m)\pi\right)\Big]\, d\nu \Bigg\}. \end{aligned} \tag{39}$$

At this point we can evaluate the geometric m-sums since $\mathrm{Im}(\nu)$ is positive in the first integral and negative in the second. We also write the integral over C_1 as a singular integral plus the half-residues at the poles (since the poles are traversed counterclockwise). The integral over C_2 is written as a singular integral minus the half-residues at the poles, and thus

$$\begin{aligned} Q_1(s,\theta) = -2\mathrm{Re}\Bigg\{ &\int_0^{i\infty} \frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \frac{e^{i\nu\theta}}{2i\sin(\nu\pi)} d\nu \\ &- \int_0^{-i\infty} \frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \frac{e^{i\nu\theta}}{2i\sin(\nu\pi)} d\nu + \pi i \sum \mathrm{Res}\left[\frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \frac{e^{i\nu\theta}}{2i\sin(\nu\pi)}\right] \\ &+ \pi i \sum \mathrm{Res}\left[\frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \frac{e^{i\nu\theta}}{2i\sin(\nu\pi)}\right] \Bigg\}. \end{aligned} \tag{40}$$

The first residue series above is over all singularities on the negative imaginary axis, and the second residue series is over all singularities on the positive imaginary axis. We can simplify (40) by making the change of variables $\nu = i\omega$ in the integrals and combining the two sums; hence

$$\begin{aligned} Q_1(s,\theta) = -2\mathrm{Re}\Bigg\{ &\int_{-\infty}^{\infty} \frac{K_{i\omega}(cs)}{K_{i\omega}(c)} I_{i\omega}(c) \frac{-ie^{-\omega\theta}}{2\sinh(\omega\pi)} d\omega \\ &+ \pi i \sum_{\nu=\pm i\omega_k} \left(\mathrm{Res}\left[\frac{K_\nu(cs)}{K_\nu(c)} I_\nu(c) \frac{e^{i\nu\theta}}{2i\sin(\nu\pi)}\right]\right) \Bigg\}. \end{aligned} \tag{41}$$

Next we will explicitly evaluate the singular integral in (41). Using $I_\nu(z) = I_{-\nu}(z) - (2/\pi)\sin(\pi\nu)K_\nu(z)$ (from [15]) with $\nu = i\omega$ yields

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{K_{i\omega}(cs)}{K_{i\omega}(c)} I_{i\omega}(c) \frac{-ie^{-\omega\theta}}{2\sinh(\omega\pi)} d\omega \\ &= -\frac{1}{2\pi}\int_{-\infty}^{\infty} K_{i\omega}(cs) e^{-\omega\theta} d\omega + \int_{-\infty}^{\infty} \frac{K_{i\omega}(cs)}{K_{i\omega}(c)} \left[I_{-i\omega}(c) + I_{i\omega}(c)\right] \frac{-ie^{-\omega\theta} d\omega}{4\sinh(\omega\pi)}. \end{aligned} \tag{42}$$

In the first integral in the right-hand side of (42) we can replace $\exp(-\omega\theta)$ by $\cosh(\omega\theta)$ and use the identities from [17],

$$\int_0^{\infty} \cos(bx)\cosh\left(\frac{x\pi}{2}\right) K_{ix}(a) dx = \frac{\pi\cos\left(a\sinh(b)\right)}{2}, \tag{43}$$

$$\int_0^\infty \sin(bx)\sinh\left(\frac{x\pi}{2}\right)K_{ix}(a)dx = \frac{\pi\sin\left(a\sinh(b)\right)}{2}, \tag{44}$$

with $a = cs,\ b = i\left(\frac{\pi}{2} - \theta\right)$, to obtain

$$\begin{aligned} &\int_{-\infty}^\infty \frac{K_{i\omega}(cs)}{K_{i\omega}(c)} I_{i\omega}(c)\frac{-ie^{-\omega\theta}}{2\sinh(\omega\pi)}d\omega \\ &= -\frac{1}{2}e^{-cs\cos\theta} + \int_{-\infty}^\infty \frac{K_{i\omega}(cs)}{K_{i\omega}(c)}\left[I_{-i\omega}(c) + I_{i\omega}(c)\right]\frac{-ie^{-\omega\theta}}{4\sinh(\omega\pi)}d\omega. \end{aligned} \tag{45}$$

Now we recall that we are interested in only the real part of (45). Since $K_{i\omega}(c)$ is real for real ω and $I_{i\omega}(c) + I_{-i\omega}(c)$ is real by the reflection principle, the integral in the right-hand side of (45) is purely imaginary, and (45) thus gives

$$\text{Re}\left\{\int_{-\infty}^\infty \frac{K_{i\omega}(cs)}{K_{i\omega}(c)} I_{i\omega}(c)\frac{-ie^{-\omega\theta}}{2\sinh(\omega\pi)}d\omega\right\} = -\frac{1}{2}e^{-cs\cos\theta}, \tag{46}$$

which when put back into (41) yields

$$Q_1(s,\theta) = e^{-cs\cos\theta} - \text{Re}\left\{2\pi i \sum_{\nu=\pm i\omega_k}\left(\text{Res}\left[\frac{K_\nu(cs)}{K_\nu(c)}I_\nu(c)\frac{e^{i\nu\theta}}{2i\sin(\nu\pi)}\right]\right)\right\}. \tag{47}$$

To compute the residues we first use the identity [15]

$$I_\nu(c)K_{\nu+1}(c) + I_{\nu+1}(c)K_\nu(c) = \frac{1}{c} \tag{48}$$

evaluated at $\nu = i\omega_k$ to replace I_ν with $(cK_{\nu+1})^{-1}$. Then computing the residues at $\nu = i\omega_k$ and $\nu = -i\omega_k$ separately, we obtain

$$\begin{aligned} Q_1(s,\theta) &= e^{-cs\cos\theta} \\ &\quad + \text{Re}\left\{\frac{\pi i}{c}\sum_{k=0}^\infty \frac{K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)}\left(\frac{e^{-\omega_k\theta}}{K_{i\omega_k+1}(c)} + \frac{e^{\omega_k\theta}}{K_{i\omega_k-1}(c)}\right)\right\}. \end{aligned} \tag{49}$$

Now, $i \times (\dot{K}_{i\omega_k}(c))^{-1}$ is real, and thus

$$\begin{aligned} Q_1(s,\theta) &= e^{-cs\cos\theta} + \frac{\pi i}{c}\sum_{k=0}^\infty \frac{K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)}\text{Re}\left(\frac{e^{-\omega_k\theta}}{K_{i\omega_k+1}(c)} + \frac{e^{\omega_k\theta}}{K_{i\omega_k-1}(c)}\right) \\ &= e^{-cs\cos\theta} + \frac{\pi i}{c}\sum_{k=0}^\infty \frac{K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)}\text{Re}\left(\frac{2\cosh(\omega_k\theta)}{K_{i\omega_k+1}(c)}\right) \\ &= e^{-cs\cos\theta} + \text{Re}\left\{\frac{\pi i}{c}\sum_{k=0}^\infty \frac{2\cosh(\omega_k\theta)K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)K_{i\omega_k+1}(c)}\right\}. \end{aligned} \tag{50}$$

Finally we combine (50), (38), (10), and (9) and use $q = (p-1)\exp(-cx)$ to obtain (13) in Theorem 2.1.

The expression in (14) follows upon using the integral representation

$$K_\nu(z) = \frac{1}{2}\int_{-\infty}^\infty e^{\nu t}e^{-z\cosh t}dt \tag{51}$$

to evaluate the integral over s as

$$\int_r^\infty e^{c(2r-s)\cos\theta}K_\nu(cs)ds = \frac{1}{2c}\int_{-\infty}^\infty \frac{e^{\nu t}e^{cr(\cos\theta-\cosh t)}}{\cos\theta+\cosh t}dt. \tag{52}$$

We mention that while (13) and (14) are exact forms of the solution, their usefulness is still numerically constrained to situations involving moderate values of c, as are the forms given in [1] and by (4), (9)–(12).

4. Shadow region asymptotics. We begin the derivation by obtaining an asymptotic formula for the roots ω_k of $K_{i\omega}(c) = 0$. We then use that formula to obtain the asymptotics for the various modified Bessel functions appearing in (13). After substituting everything into (13) we drop all but the first term of the k-series and use Laplace's method to expand the s-integral. The terms with $k \geq 1$ in (13) lead to exponentially small errors in the shadow region.

To obtain the asymptotics of ω_k we first apply the saddle point method to (51). We set $\nu = i\omega_k = ic\beta$ and obtain the following for $0 < \beta < 1$:

$$K_{ic\beta}(c) \sim \sqrt{\frac{\pi}{2c}}\left(1-\beta^2\right)^{-1/4}\exp\left[-c\left(\beta\sin^{-1}(\beta)+\sqrt{1-\beta^2}\right)\right]. \tag{53}$$

For $\beta \approx 1$, specifically $\beta = 1+O(c^{-2/3})$, (53) becomes invalid. This range corresponds to two saddles coalescing, and then we need to approximate $K_{i\beta c}(c)$ by Airy functions as follows:

$$\begin{aligned} K_{i\left(c+c^{1/3}\alpha\right)}(c) &\sim \pi\left(\frac{2}{c}\right)^{1/3}\exp\left[-\frac{\pi}{2}\left(c+c^{1/3}\alpha\right)\right]\Bigg\{\mathrm{Ai}(-2^{1/3}\alpha) \\ &+ c^{-2/3}\left[\frac{-2\alpha}{15}\mathrm{Ai}(-2^{1/3}\alpha)+\frac{2^{1/3}\alpha^2}{30}\mathrm{Ai}'(-2^{1/3}\alpha)\right] \\ &+ c^{-4/3}\left[2^{1/3}\left(-\frac{\alpha^3}{105}+\frac{1}{70}\right)\mathrm{Ai}'(-2^{1/3}\alpha)+\left(-\frac{\alpha^5}{1800}+\frac{\alpha^2}{25}\right)\mathrm{Ai}(-2^{1/3}\alpha)\right]\Bigg\}. \end{aligned} \tag{54}$$

If the right-hand side of (54) is to vanish, we need $\alpha \sim -2^{-1/3}r_k$ so that $\omega_k \sim c + c^{1/3}2^{-1/3}|r_k|$. To find the next term we expand the Airy functions in (54) near $\alpha = -2^{-1/3}r_k$ and set $\alpha - 2^{-1/3}|r_k| \sim c^{-2/3}\gamma_k$ to find that $\gamma_k = 2^{-2/3}r_k^2/30$. We then repeat the process with $\alpha - 2^{-1/3}|r_k| \sim c^{-2/3}\gamma_k + c^{-4/3}\beta_k$ to find that $\beta_k = (r_k^3+10)/700$; thus

$$\omega_k = c + c^{1/3}2^{-1/3}|r_k| + c^{-1/3}\frac{2^{-2/3}r_k^2}{30} + c^{-1}\frac{r_k^3+10}{700} + O\left(c^{-5/3}\right). \tag{55}$$

Now that we have the expansion of ω_k we can derive the expansions for the modified Bessel functions in (13). We use (55) in (51) with $z = c$ and $\nu = i\omega_k$ or $\nu = i\omega_k + 1$ and expand the integral by the saddle point method, noting that there are two saddles near $t = i\pi/2$ and using also the derivatives of the Airy function, in the form

$$\mathrm{Ai}^{(n)}(z) = \frac{1}{2\pi i}\int_C v^n e^{zv-v^3/3}dv. \tag{56}$$

Here the integration contour runs from $\infty e^{-2\pi i/3}$ to the origin and then to $\infty e^{2\pi i/3}$ (C could also be taken as the imaginary axis in the ν-plane). After some calculation we obtain

$$\begin{aligned}\dot{K}_{i\omega_k}(c) = c^{-2/3}2^{2/3}\pi i\mathrm{Ai}'(r_k)\exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right]\\ \times\left[1-c^{-1/3}\frac{\pi\gamma_k}{2}+c^{-2/3}\left(\frac{\pi^2\gamma_k^2}{8}-\frac{A_k}{5}\right)+O\left(c^{-1}\right)\right]\end{aligned}\tag{57}$$

and

$$\begin{aligned}K_{i\omega_k+1}(c) \sim -c^{-2/3}2^{2/3}\pi\mathrm{Ai}'(r_k)\exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right]\\ \times\left[1-c^{-1/3}\frac{\pi\gamma_k}{2}+c^{-2/3}\left(\frac{\pi^2\gamma_k^2}{8}-\frac{2A_k}{15}\right)+c^{-1}\left(-\frac{A_k\gamma_k\pi}{15}-\frac{\beta_k\pi}{2}-\frac{\gamma_k^3\pi^3}{48}\right)\right].\end{aligned}\tag{58}$$

Here $A_k = 2^{-1/3}|r_k|$, where r_k satisfy $\mathrm{Ai}(r_k) = 0$ with $0 > r_0 > r_1 > r_2 > \cdots$.

Next we consider (51) with $z = cs$ and note that $s > r$ in the s-integral in (13), and that $r > 1$ in the shadow region. Then we obtain approximations to $K_{i\omega_k}(cs)$ and $K_{i\omega_k+1}(cs)$ by again using the saddle point method on (51), but now a single saddle at $t = i\sin^{-1}(1/s)$ (thus $|t| < \pi/2$) determines the asymptotic behavior of the integral(s). After a lengthy calculation we obtain

$$\begin{aligned}K_{i\omega_k}(cs) = c^{-1/2}\sqrt{\frac{\pi}{2}}\left(s^2-1\right)^{-1/4}\exp\left[-\sin^{-1}\left(\frac{1}{s}\right)\left(c+c^{-1/3}A_k\right)\right]\\ \times\exp\left[-c\sqrt{s^2-1}\right]\left\{1+c^{-1/3}\left[\frac{-A_k^2}{2\left(s^2-1\right)^{1/2}}-\gamma_k\sin^{-1}\left(\frac{1}{s}\right)\right]\right.\\ +c^{-2/3}\left[\frac{\gamma_k^2\left(\sin^{-1}\left(1/s\right)\right)^2}{2}+\frac{A_k^4+4A_k}{8\left(s^2-1\right)}+\frac{\gamma_kA_k^2\sin^{-1}\left(1/s\right)}{2\left(s^2-1\right)^{\frac{1}{2}}}\right]\\ +c^{-1}\left[\frac{-\gamma_k^3\left(\sin^{-1}\left(1/s\right)\right)^3}{6}-\frac{A_k\gamma_k\sin^{-1}\left(1/s\right)}{5}-\frac{A_k^6}{48\left(s^2-1\right)^{3/2}}\right.\\ -\frac{5}{24\left(s^2-1\right)^{3/2}}-\frac{1}{8\left(s^2-1\right)^{1/2}}-\frac{A_k\gamma_k}{\left(s^2-1\right)^{1/2}}-\frac{5A_k^3}{12\left(s^2-1\right)^{3/2}}\\ \left.\left.-\frac{\left(A_k^4+4A_k\right)\gamma_k\sin^{-1}\left(1/s\right)}{8\left(s^2-1\right)}-\frac{A_k^2\gamma_k^2\left(\sin^{-1}\left(1/s\right)\right)^2}{4\left(s^2-1\right)^{1/2}}\right]+O\left(c^{-4/3}\right)\right\}\end{aligned}\tag{59}$$

and

$$\begin{aligned}K_{i\omega_k+1}(cs) = c^{-1/2}\sqrt{\frac{\pi}{2}}\left(s^2-1\right)^{-\frac{1}{4}}\exp\left[-\sin^{-1}\left(\frac{1}{s}\right)\left(c+c^{-1/3}A_k-i\right)\right]\\ \times\exp\left[-c\sqrt{s^2-1}\right]\left\{1+c^{-1/3}\left[\frac{-A_k^2}{2\left(s^2-1\right)^{1/2}}-\gamma_k\sin^{-1}\left(\frac{1}{s}\right)\right]\right.\\ +c^{-2/3}\left[\frac{\gamma_k^2\left(\sin^{-1}\left(1/s\right)\right)^2}{2}+\frac{A_k^4+4A_k}{8\left(s^2-1\right)}+\frac{\gamma_kA_k^2\sin^{-1}\left(1/s\right)}{2\left(s^2-1\right)^{1/2}}+\frac{iA_k}{\sqrt{s^2-1}}\right]\\ +c^{-1}\left[\frac{-i\left(A_k^3+1\right)}{2\left(s^2-1\right)}-\frac{iA_k\gamma_k\sin^{-1}\left(1/s\right)}{\left(s^2-1\right)^{1/2}}\right.\end{aligned}$$

$$
\begin{aligned}
&- \frac{\gamma_k^3\left(\sin^{-1}(1/s)\right)^3}{6} - \frac{A_k\gamma_k\sin^{-1}(1/s)}{5} - \frac{A_k^6}{48\left(s^2-1\right)^{3/2}} \\
&- \frac{5}{24\left(s^2-1\right)^{3/2}} + \frac{3}{8\left(s^2-1\right)^{1/2}} - \frac{A_k\gamma_k}{\left(s^2-1\right)^{1/2}} - \frac{5A_k^3}{12\left(s^2-1\right)^{3/2}} \\
&\left. - \frac{\left(A_k^4+4A_k\right)\gamma_k\sin^{-1}(1/s)}{8\left(s^2-1\right)} - \frac{A_k^2\gamma_k^2\left(\sin^{-1}(1/s)\right)^2}{4\left(s^2-1\right)^{1/2}}\right] + O\left(c^{-4/3}\right)\left.\vphantom{\frac{1}{1}}\right\}. \qquad (60)
\end{aligned}
$$

Computing the reciprocals of (57) and (58) and using these along with (59) and (60) in (13), we obtain

$$
\begin{aligned}
p = c^{11/6}&\int_r^\infty e^{c(\cos\theta)(2r-s)}\sum_{k=0}^\infty \exp\left[\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)\left(c+c^{1/3}A_k\right)\right] \\
&\times \frac{2^{1/6}\left(1-s\sin\theta\right)}{\sqrt{\pi}\left[\mathrm{Ai}'(r_k)\right]^2 s\left(s^2-1\right)^{1/4}}\exp\left[-c\sqrt{s^2-1}\right] \qquad (61)\\
&\times\left\{1+c^{-1/3}\left[\gamma_k\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)-\frac{A_k^2}{2\left(s^2-1\right)^{1/2}}\right)\right]+O(c^{-2/3})\right\}ds.
\end{aligned}
$$

We approximate (61) by retaining only the $k=0$ term, since it is exponentially larger (by a factor $\exp[O(c^{1/3})]$) than the terms with $k\geq 1$. The result (15) follows upon using Laplace's method to expand the integral. The main contribution comes from near the lower limit of integration $s=r$.

5. Shadow boundary asymptotics. For the intermediate shadow boundary region, where $y-1=O(c^{-1/3})$ and $x>0$, we begin similarly to the derivation of Theorem 2.2. But now we will keep the entire k-series in (13), not just the $k=0$ term. After substituting (59) and (60) and the reciprocals of (57) and (58) into (13) and determining several terms in the expansion of the resulting s-integral using Laplace's method, we are led to

$$
\begin{aligned}
p = \sum_{k=0}^\infty &\frac{c^{5/6}2^{1/6}}{\sqrt{\pi}\left[\mathrm{Ai}'(r_k)\right]^2}\exp\left[c\left(r\cos\theta-\sqrt{r^2-1}+\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)\right] \\
&\times\left(r^2-1\right)^{-1/4}\frac{1-r\sin\theta}{r\cos\theta+\sqrt{r^2-1}}\exp\left[c^{1/3}A_k\left(\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)\right] \\
&\times\left\{1+c^{-1/3}\left[\gamma_k\left(\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)-\frac{A_k^2}{2\sqrt{r^2-1}}\right]\right. \\
&+c^{-2/3}\left[\frac{A_k\left(r^2-1\right)^{-1/2}}{r\cos\theta+\sqrt{r^2-1}}+\frac{16A_k}{15}+\frac{A_k^4+4A_k}{8\left(r^2-1\right)}+\frac{A_k}{1-r\sin\theta}\right. \\
&\qquad\left.+\frac{\gamma_k^2\left[\theta-\sin^{-1}(1/r)\right]^2}{2}+\frac{\gamma_kA_k^2\left[\theta-\sin^{-1}(1/r)\right]}{2\sqrt{r^2-1}}\right] \\
&+c^{-1}\frac{1}{1-r\sin\theta}\left[A_k\gamma_k\left[\theta-\sin^{-1}\left(\frac{1}{r}\right)\right]-\frac{A_k^3}{2\sqrt{r^2-1}}\right. \\
&\qquad\left.\left.-\frac{1}{r\cos\theta+\sqrt{r^2-1}}\right]+O(c^{-1})\right\}. \qquad (62)
\end{aligned}
$$

Note that we have written in (62) only a part of the $O(c^{-1})$ term inside the $\{\cdot\}$. This is the part that contains the factor $(1 - r\sin\theta)^{-1} = (1-y)^{-1}$, which becomes large as $y \to 1$.

In the shadow boundary we have $1 - r\sin\theta = 1 - y = c^{-1/3}Y$ and we rewrite (62) in terms of the variables (x, Y), also using the expansions

$$\sqrt{r^2-1} = x + c^{-1/3}\frac{Y}{x} + c^{-2/3}Y^2\frac{x^2-1}{2x^3} + c^{-1}Y^3\frac{1-x^2}{2x^5} + O(c^{-4/3}), \tag{63}$$

$$\theta - \sin^{-1}\left(\frac{1}{r}\right) = c^{-1/3}\frac{Y}{x} - c^{-2/3}\frac{Y^2}{2x^3} + c^{-1}Y^3\frac{3-2x^2}{6x^5} + O(c^{-4/3}), \tag{64}$$

$$\frac{1 - r\sin\theta}{r\cos\theta + \sqrt{r^2-1}} = -c^{-1/3}\frac{Y}{2x} + c^{-2/3}\frac{Y^2}{4x^3} + c^{-1}Y^3\frac{x^2-2}{8x^5} + O(c^{-4/3}). \tag{65}$$

After substituting (63)–(65) into (62) and simplifying, we obtain (17)–(20). But, this formula is valid only for $Y < 0$ since the series in (17) will not converge for $Y > 0$.

To transform (17)–(20) into the integral expressions in (21)–(24) we first observe that the residues at $z = -A_k$ are

$$\operatorname{Res}\left[e^{-zY/x}\frac{z^n}{\left(\operatorname{Ai}(2^{1/3}z)\right)^2}\right]_{z=-A_k} = \frac{2^{-2/3}}{\left[\operatorname{Ai}'(r_k)\right]^2}e^{A_kY/x}\left[n\left(-A_k\right)^{n-1} - \left(-A_k\right)^n\frac{Y}{x}\right], \tag{66}$$

which we use to obtain the identity

$$\begin{aligned}&\frac{2^{2/3}}{2\pi i}\int_C \left(a_0 + a_1z + a_2z^2 + a_3z^3 + a_4z^4\right)e^{-zY/x}\frac{dz}{\left[\operatorname{Ai}(2^{1/3}z)\right]^2}\\ &= \sum_{k=0}^{\infty}\frac{e^{A_kY/x}}{\left[\operatorname{Ai}'(r_k)\right]^2}\left[a_1 - a_0\frac{Y}{x} - A_k\left(2a_2 - a_1\frac{Y}{x}\right)\right.\\ &\qquad \left. + A_k^2\left(3a_3 - a_2\frac{Y}{x}\right) - A_k^3\left(4a_4 - a_3\frac{Y}{x}\right) - A_k^4a_4\frac{Y}{x}\right],\end{aligned} \tag{67}$$

where $a_i \in \mathbf{R}$ and the integration contour is the imaginary axis.

Using (67) with $a_0 = x^{-1/2}$ and $a_1 = a_2 = a_3 = a_4 = 0$, we see the equivalence of the leading terms in (17) and (21) (with (18) and (22)). A similar calculation shows the equivalence of the second terms in (17) and (21) (with (19) and (23)). In view of the rapid decay of $\operatorname{Ai}^{-2}(2^{1/3}z)$ in the imaginary directions ($z \to \pm i\infty$), the integrals in (22) and (23) converge for all real Y and give the continuation of the sums in (17) to the range $Y > 0$ ($y > 1$).

We next attempt to write R_2 in (20) as a similar integral and show that this is not quite possible. Since (20) is quartic in the A_k, we would need a fourth order polynomial in z as in (67). By examining (20) and writing it in the form of the right-hand side of (67), we would need

$$a_0 = \frac{Y^8}{x^8}\frac{\left(x^2-1\right)^2}{128x^{5/2}} + \frac{Y^5}{x^5}\frac{15-19x^2}{80x^{5/2}} + \frac{Y^2}{x^2}\frac{15-4x^2}{24x^{5/2}} - \frac{x^{1/2}}{Y}. \tag{68}$$

The last $x^{1/2}/Y$ term in (68) is problematic and occurs as a consequence of a "missing" term in (20). From the derivation of (17)–(20), we see that all terms appearing in (18)–(20) arise from the Q_2 contribution to (13) (cf. (38)). Had we considered only the Q_2 contribution in the derivation of (18)–(20), one other term, $-x^{-1/2}$, would also appear in (20). It is absent because it cancels with a contribution from Q_1 (the only other contribution of Q_1 to (17)–(20) is the cancellation of the -1 in $p-1$). The part of (20) that arises from Q_2 (including the "missing" term, $-x^{-1/2}$) can be represented in integral form, using (67) with

$$a_0 = \frac{Y^8}{x^8}\frac{\left(x^2-1\right)^2}{128x^{5/2}} + \frac{Y^5}{x^5}\frac{15-19x^2}{80x^{5/2}} + \frac{Y^2}{x^2}\frac{15-4x^2}{24x^{5/2}}, \tag{69}$$

$$a_1 = \frac{Y^6}{x^6}\frac{x^2-1}{16x^{5/2}} + \frac{Y^3}{x^3}\frac{4x^2-9}{12x^{5/2}} - \frac{1+2x^2}{2x^{5/2}}, \tag{70}$$

$$a_2 = \frac{Y^4}{x^4}\frac{3-x^2}{16x^{5/2}} + \frac{Y}{x}\frac{2x^2+45}{60x^{5/2}}, \tag{71}$$

$$a_3 = -\frac{Y^2}{x^2}\frac{1}{4x^{5/2}}, \tag{72}$$

$$a_4 = \frac{1}{8x^{5/2}}, \tag{73}$$

which gives us the integral portion of (24). Adding to it the contribution from Q_1 (which is $c^{-2/3}x^{-1/2}$ times the leading multiplicative factors in (17)) produces (24).

The first two terms in (21) can be extended for $Y > 0$, and when expanded for $Y \to +\infty$ they match to the asymptotic results for $p-1$ in the illuminated region in [2] when the latter are expanded for $y \downarrow 1$. They also asymptotically match to the results for $p-1$ at the birth of the shadow boundary where $x \approx 0$ and $y \approx 1$, which are also given in [2].

To see the appearance of the uniform concentration at infinity ($= 1$), we need to consider $Y \approx 0$ or, more precisely, $y - 1 = O(c^{-1/2})$, which corresponds to $Y = O(c^{-1/6})$. We thus set $y - 1 = c^{-1/2}\Delta$ or $Y = c^{-1/6}\Delta$ and examine $p-1$ as given by (4) with (9)–(12). We will examine separately the contributions from Q_1 in (11) and Q_2 in (12). We define $Q1$ and $Q2$, which are functions of (r,θ) or (x,y), by

$$Q1 = -2c\cos\theta \int_r^\infty e^{c(2r-s)\cos\theta} Q_1(s,\theta)ds \tag{74}$$

and

$$Q2 = \int_r^\infty e^{c(2r-s)\cos\theta} Q_2(s,\theta)ds \tag{75}$$

and note that $p-1$ is equal to the sum of (74) and (75).

The expansion of $Q2$ on the Δ-scale can be obtained as a limiting form of the integral expressions in (22)–(24), to four orders in $c^{-1/6}$. We begin by noting that $\mathcal{F}_0$ in (22) and $\mathcal{F}_1$ in (23) are smooth functions of Y and can be expanded as $Y =$

$c^{-1/6}\Delta \to 0$ using the Taylor expansion of $\exp[-zY/x] = \exp[-c^{-1/6}\Delta z/x]$. This leads to

$$\begin{aligned}\mathcal{F}_0 = 2^{-1/3}x^{-1/2}\Bigg[&1 - c^{-1/6}2^{-1/3}C_1\frac{\Delta}{x}\\ &+ c^{-1/3}2^{-2/3}C_2\frac{\Delta^2}{2x^2} - c^{-1/2}C_3\frac{\Delta^3}{12x^3} + O(c^{-2/3})\Bigg]\end{aligned} \tag{76}$$

and

$$\mathcal{F}_1 = -x^{-3/2}\left[\frac{C_2}{4} + c^{-1/6}2^{-1/3}\frac{\Delta\,(2-C_3)}{4x} + O(c^{-1/3})\right]. \tag{77}$$

The constants C_i above are defined in (30). The integral portion of (24) (which arose from $Q2$) remains $O(1)$ as $Y \to 0$ and will not contribute to (25) until the fifth order. We now change variables in (21) from Y to Δ and expand the exponential $\exp\left[c^{-1/2}\Delta^3/(6x^3)\right]$. Using (76) and (77), we thus obtain

$$\begin{aligned}Q2 &\sim c^{1/2}2^{-1/6}\pi^{-1/2}\exp\left[-c^{1/3}\frac{Y^2}{2x}\right]\exp\left[\frac{Y^3}{6x^3}\right]\left\{\mathcal{F}_0 + c^{-1/3}\mathcal{F}_1\right\}\\ &= \frac{c^{1/2}}{\sqrt{2\pi x}}\exp\left(-\frac{\Delta^2}{2x}\right)\Bigg\{1 - c^{-1/6}\left(\frac{2^{-1/3}C_1}{\sqrt{x}}\right)\frac{\Delta}{\sqrt{x}}\\ &\quad + c^{-1/3}\left(\frac{2^{-2/3}C_2}{2x}\right)\left[\left(\frac{\Delta}{\sqrt{x}}\right)^2 - 1\right]\\ &\quad + c^{-1/2}\left(\frac{2-C_3}{12x\sqrt{x}}\right)\left[\left(\frac{\Delta}{\sqrt{x}}\right)^3 - 3\frac{\Delta}{\sqrt{x}}\right] + O(c^{-2/3})\Bigg\}.\end{aligned} \tag{78}$$

We see that the expansion of $Q2$ involves the Hermite polynomials in the similarity variable $\Delta/\sqrt{x}$.

Next we consider $Q1$ on the Δ-scale. Substituting (50) into (74) yields

$$Q1 = -1 + \mathrm{Im}\left[\int_r^\infty e^{c(2r-s)\cos\theta}\sum_{k=0}^{\infty}\frac{4\pi\cos\theta\cosh(\omega_k\theta)K_{i\omega_k}(cs)}{\sinh(\omega_k\pi)\dot{K}_{i\omega_k}(c)K_{i\omega_k+1}(c)}ds\right]. \tag{79}$$

In the derivation of the shadow region asymptotics we needed to consider only the $k = 0$ term of the series. In the shadow boundary intermediate region (cf. (17)) we needed to keep all terms of the series, and this led to the first term in the right-hand side of (24) for $Y < 0$. On the Δ-scale, not only will we have to keep all terms of the k-series, but the terms with k large will actually play a larger role. Our previous expansions of the modified Bessel functions in (57)–(59) were derived under the assumption $r_k = O(1)$ and break down when $r_k = O(c^{1/6})$, since then the second terms become of the same order as the first terms. Since $r_k = O(k^{2/3})$ for $k \to \infty$ [15], this occurs when $k = O(c^{1/4})$, so our previous expansions are valid only for the $k \ll c^{1/4}$ terms of the series. We redo the saddle point analysis used to obtain (57)–(59), this time considering $r_k = O(c^{1/6})$. This yields to leading order

$$\dot{K}_{i\omega_k}(c) \sim c^{-2/3}2^{2/3}\pi i\mathrm{Ai}'(r_k)\exp\left[-\frac{\pi}{2}\left(c + c^{1/3}A_k + c^{-1/3}\gamma_k\right)\right], \tag{80}$$

$$K_{i\omega_k+1}(c) \sim -c^{-2/3}2^{2/3}\pi\mathrm{Ai}'(r_k)\exp\left[-\frac{\pi}{2}\left(c+c^{1/3}A_k+c^{-1/3}\gamma_k\right)\right], \tag{81}$$

and

$$\begin{aligned} K_{i\omega_k}(cs) &\sim c^{-1/2}\left(s^2-1\right)^{-1/4}\sqrt{\frac{\pi}{2}}\exp\left[-2^{-5/3}\left(s^2-1\right)^{-1/2}\frac{r_k^2}{c^{1/3}}\right] \\ &\times\exp\left[-c\sqrt{s^2-1}\right]\exp\left[-\sin^{-1}\left(\frac{1}{s}\right)\left(c+c^{1/3}A_k+c^{-1/3}\gamma_k\right)\right]. \end{aligned} \tag{82}$$

Note that $\gamma_k = O(c^{1/3})$ for $r_k = O(c^{1/6})$. Substituting (80)–(82) into (79) gives

$$\begin{aligned} Q1 \sim -1 &+ \int_r^\infty \exp\left[c\left((2r-s)\cos\theta-\sqrt{s^2-1}+\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)\right] \\ &\times\frac{c^{5/6}2^{1/6}\cos\theta}{\sqrt{\pi}\left(s^2-1\right)^{1/4}}\sum_{k=0}^{\infty}\left\{\exp\left[-c^{1/3}2^{-1/3}r_k\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)\right]\right. \\ &\left.\times\exp\left[-\frac{r_k^2}{4}\left(c^{-1/3}2^{1/3}\left(s^2-1\right)^{-1/2}\right)\right]\left[\mathrm{Ai}'(r_k)\right]^{-2}\right\}ds. \end{aligned} \tag{83}$$

We now replace the k-series with an integral. To do so, we will utilize an identity derived in [16],

$$\begin{aligned} &\int_0^\infty \mathrm{Ai}(\tau+M)e^{N\tau}d\tau + 2\mathrm{Re}\left[\int_0^\infty\frac{\mathrm{Ai}(\omega\tau+M)}{\mathrm{Ai}(\omega\tau)}\mathrm{Ai}(\tau)e^{\omega N\tau}d\tau\right] \\ &= e^{-MN}e^{N^3/3}-\sum_{k=0}^{\infty}e^{Nr_k}\frac{\mathrm{Ai}(r_k+M)}{\left[\mathrm{Ai}'(r_k)\right]^2}, \end{aligned} \tag{84}$$

where $\omega = \exp[2\pi i/3]$. Note that the series in the right-hand side of (84) converges only for $N > 0$, but the integrals in the left-hand side converge for all N. First we expand the Airy functions for large M using the well-known expansion [15]

$$\mathrm{Ai}(\tau+M) \sim \frac{1}{2\sqrt{\pi}}M^{-1/4}\exp\left[-\frac{2}{3}M^{3/2}\right]\exp\left[-\tau\sqrt{M}-\frac{\tau^2}{4\sqrt{M}}\right]. \tag{85}$$

For $M \to \infty$ the second integral in (84) will be asymptotically negligible compared to the first. After replacing the Airy functions with their expansions we use the exact relationship $-MN + N^3/3 = -2M^{3/2}/3 + M^{1/2}(N-M^{1/2})^2 + (N-M^{1/2})^3/3$ and make the change of variables $\tau = 2^{1/4}M^{1/4}q$ in the first integral in (84). We then cancel the common factors $\exp\left[-2M^{3/2}/3\right]$ and rearrange terms to obtain

$$\begin{aligned} &\frac{2^{1/4}}{2\sqrt{\pi}}\int_0^\infty \exp\left[2^{1/4}M^{1/4}\left(N-M^{1/2}\right)q\right]\exp\left[-\frac{q^2}{2\sqrt{2}}\right]dq \\ &\qquad -\exp\left[M^{1/2}\left(N-M^{1/2}\right)^2+\frac{1}{3}\left(N-M^{1/2}\right)^3\right] \\ &\sim -\frac{M^{-1/4}}{2\sqrt{\pi}}\sum_{k=0}^{\infty}\frac{\exp\left[r_k\left(N-M^{1/2}\right)\right]\exp\left[-\frac{r_k^2}{4}M^{-1/2}\right]}{\left[\mathrm{Ai}'(r_k)\right]^2}. \end{aligned} \tag{86}$$

This result applies for $N,\ M \to \infty$ with $M^{1/4}\left(N-M^{1/2}\right) = O(1)$. To apply (86) to the k-sum in (83), we set $M = 2^{-2/3}c^{2/3}\left(s^2-1\right)$ and $N = 2^{-1/3}c^{1/3}[\left(s^2-1\right)^{1/2} -$

$\left(\theta - \sin^{-1}(1/s)\right)]$, which yields

$$\begin{aligned}
&\sum_{k=0}^{\infty} \frac{\exp\left[-r_k\left(2^{-1/3}c^{1/3}\right)\left(\theta-\sin^{-1}(1/s)\right)\right]\exp\left[-c^{-1/3}2^{-5/3}r_k^2\left(s^2-1\right)^{-1/2}\right]}{\left[\mathrm{Ai}'(r_k)\right]^2} \\
&\quad\sim\; c^{1/6}2^{5/6}\sqrt{\pi}\left(s^2-1\right)^{1/4}\exp\left[\frac{c}{2}\sqrt{s^2-1}\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)^2\right] \\
&\times\exp\left[-\frac{c}{6}\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)^3\right]-c^{1/6}2^{1/12}\left(s^2-1\right)^{1/4}\int_0^{\infty}\exp\left[-\frac{q^2}{2\sqrt{2}}\right] \\
&\qquad\times\exp\left[-c^{1/2}2^{-1/4}\left(s^2-1\right)^{1/4}\left(\theta-\sin^{-1}\left(\frac{1}{s}\right)\right)q\right]dq.
\end{aligned}\tag{87}$$

After substituting (87) into (83) and asymptotically evaluating the integral over s, we have

$$\begin{aligned}
Q1 \sim\; &-1+\frac{2r\cos\theta}{r\cos\theta+\sqrt{r^2-1}}\exp\!\left[c\left(r\cos\theta-\sqrt{r^2-1}+\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)\right] \\
&\times\left\{\exp\left[\frac{c}{2}\sqrt{r^2-1}\left(\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)^2-\frac{c}{6}\left(\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)^3\right]\right. \\
&\left.-\frac{2^{1/4}}{2\sqrt{\pi}}\int_0^{\infty}\exp\left[-c^{1/2}2^{-1/4}\left(r^2-1\right)^{1/4}\left(\theta-\sin^{-1}\left(\frac{1}{r}\right)\right)q-\frac{q^2}{2\sqrt{2}}\right]dq\right\}.
\end{aligned}\tag{88}$$

Finally we set $Y = c^{-1/6}\Delta$ in (63) and (64) and rewrite (88) in terms of (x, Δ). To leading order this yields

$$Q1 \sim -\frac{2^{1/4}}{2\sqrt{\pi}}\exp\left[-\frac{\Delta^2}{2x}\right]\int_0^{\infty}\exp\left[-2^{-1/4}q\frac{\Delta}{\sqrt{x}}-\frac{q^2}{2\sqrt{2}}\right]dq.\tag{89}$$

Substituting (89) and (78) into $p = 1 + Q1 + Q2$ gives the result in (25)–(29). The error function in (89) reveals how the concentration p transitions from being small in the shadow region to being approximately unity in the illuminated region.

6. Shadow part of obstacle boundary. The derivation of the asymptotics for the nested layers close to the obstacle on the shadow side is similar to the derivation of the outer shadow region from Theorem 2.2. The main difference is that we will have to reexamine the expansions of the modified Bessel functions involving s. Our previous expansions (59) and (60) break down as $s \to 1$.

For the intermediate layer near the obstacle we again use (51) but make the change of variables

$$s = 1 + c^{-2/3}\rho\tag{90}$$

and approximate the Bessel functions by Airy functions, as

$$\begin{aligned}
K_{i\omega_k}(cs) = c^{-1/3}2^{1/3}\pi\mathrm{Ai}\left(\widehat{\rho}_k\right)\exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right] \\
\times\left\{1-c^{-1/3}\frac{\pi\gamma_k}{2}+O\left(c^{-2/3}\right)\right\}
\end{aligned}\tag{91}$$

and

$$\begin{aligned} K_{i\omega_k+1}(cs) &= c^{-1/3}2^{1/3}\pi i \ \exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right] \\ &\times\left\{\mathrm{Ai}\left(\widehat{\rho}_k\right)+c^{-1/3}\left(2^{1/3}i\mathrm{Ai}'\left(\widehat{\rho}_k\right)-\frac{\pi\gamma_k}{2}\mathrm{Ai}\left(\widehat{\rho}_k\right)\right)+O(c^{-2/3})\right\}, \end{aligned}\tag{92}$$

where $\widehat{\rho}_k = 2^{1/3}(\rho - A_k)$. We use (91), (92) and the reciprocals of (57) and (58) in (13). Only the $k = 0$ term in the series contributes, and after simplification we get

$$\begin{aligned} p &= c^{4/3}\frac{2\left(1-\sin\theta\right)}{\left[\mathrm{Ai}'(r_0)\right]^2}\exp\left[c\left(2r-1\right)\left(\cos\theta\right)+\left(c+c^{1/3}A_0\right)\left(\theta-\frac{\pi}{2}\right)\right] \\ &\times\int_{c^{2/3}(r-1)}^{\infty}e^{-c^{1/3}\rho\cos\theta}\mathrm{Ai}\left[2^{1/3}\left(\rho-A_0\right)\right]\left\{1+c^{-1/3}\gamma_0\left(\theta-\frac{\pi}{2}\right)+O(c^{-2/3})\right\}d\rho. \end{aligned}\tag{93}$$

We use Laplace's method to expand the integral in (93). We set $\eta = c^{2/3}\left(r-1\right) = O(1)$ (thus $r - 1 = O(c^{-2/3})$), and then the major contribution comes from the lower limit $\rho = \eta$. Setting $u = \rho - \eta$ and expanding for $u \to 0$ gives

$$\begin{aligned} p = c^{4/3}&\exp\left[c\left(2r-1\right)\cos\theta+\left(c+c^{1/3}A_0\right)\left(\theta-\frac{\pi}{2}\right)-c^{1/3}\eta\cos\theta\right] \\ &\times\ \frac{2\left(1-\sin\theta\right)}{\left[\mathrm{Ai}'(r_0)\right]^2}\int_0^{\infty}\exp\left[-c^{1/3}u\cos\theta\right] \\ &\times\ \left\{\mathrm{Ai}\left[2^{1/3}\left(\eta-A_0\right)\right]+2^{1/3}\mathrm{Ai}'\left[2^{1/3}\left(\eta-A_0\right)\right]u+O(u^2)\right\} \\ &\times\ \left[1+c^{-1/3}\gamma_0\left(\theta-\frac{\pi}{2}\right)+O(c^{-2/3})\right]du, \end{aligned}\tag{94}$$

and this leads to (31).

For the inner layer near the obstacle we use (51) with the change of variables

$$s = 1 + c^{-1}\sigma\tag{95}$$

to obtain

$$\begin{aligned} K_{i\omega_k}(cs) &= c^{-2/3}2^{2/3}\pi\sigma\mathrm{Ai}'\left(r_k\right)\exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right] \\ &\times\left\{1-c^{-1/3}\frac{\pi\gamma_k}{2}+O\left(c^{-2/3}\right)\right\} \end{aligned}\tag{96}$$

and

$$\begin{aligned} K_{i\omega_k+1}(cs) &= c^{-2/3}2^{2/3}\pi i\left(\sigma+i\right)\mathrm{Ai}'\left(r_k\right)\exp\left[-\frac{\pi}{2}\left(c+c^{-1/3}A_k\right)\right] \\ &\times\left\{1-c^{-1/3}\frac{\pi\gamma_k}{2}+O\left(c^{-2/3}\right)\right\}. \end{aligned}\tag{97}$$

We use (96), (97) and the reciprocals of (57) and (58) in (13), retain only the $k = 0$ term, and set $\xi = c\left(r-1\right) = O(1)$ to obtain

$$\begin{aligned} p &= c^{2/3}\frac{2^{4/3}\left(1-\sin\theta\right)}{\mathrm{Ai}'(r_0)}\exp\left[c\left(2r-1\right)\cos\theta+\left(c+c^{1/3}A_0\right)\left(\theta-\frac{\pi}{2}\right)\right] \\ &\times\int_{\xi}^{\infty}\sigma e^{-\sigma\cos\theta}\left[1+c^{-1/3}\gamma_0\left(\theta-\frac{\pi}{2}\right)+O(c^{-2/3})\right]d\sigma. \end{aligned}\tag{98}$$

Evaluating the integral leads to (33).

7. Conclusion. To summarize, we have obtained alternate representations (cf. (13) and (14)) of the concentration profile of some substance that undergoes drift–diffusion past a circular obstacle. From the new formulas we obtained detailed asymptotic results for the concentration in the shadow region, including the shadow boundary and the shadow side of the obstacle. These asymptotic results do not seem to be obtainable from the previous forms of the solution in [1] and [2]. They show that in the shadow region the concentration is exponentially small in c, the parameter measuring the ratio of drift to diffusion, with additional factors that vary exponentially in $c^{1/3}$ and algebraically in c. The smallness is due mainly to the term $r\cos\theta + \theta - \arcsin(1/r) - \sqrt{r^2-1} < 0$ in the exponent in (15). The concentration remains small if $r \approx 1$ (cf. Theorem 2.4) as long as $|\theta| < \pi/2$. In the shadow boundaries, where $y \approx \pm 1$ with $x > 0$, the concentration profile undergoes the transition from $p \approx 1$ to $p \approx 0$. This is governed by the two nested layers in Theorem 2.3. For the coarser spatial scale $y - (\pm 1) = O(c^{-1/3})$ the transition has already taken place and the solution involves Airy functions. For the finer spatial scale $y - (\pm 1) = O(c^{-1/2})$ the transition from $p \approx 1$ to $p \approx 0$ occurs via the error function that arises as a part of the fourth term in the asymptotic expansion. The leading term on this scale shows (cf. (26)) a Gaussian particle profile and a large $O(\sqrt{c})$ concentration, with the profile diffusing further with increasing x. For sufficiently large x the shadow region disappears, as we must have $p \to 1$ as $r \to \infty$ in any direction.

The structure of the shadow boundary is much different from corresponding scattering problems, where an incoming plane wave disappears in the region Ω_s via a Fresnel integral at the leading asymptotic order.

While exact solutions can be obtained only for simple geometries such as circles or spheres, we are currently analyzing such problems directly, via singular perturbation techniques such as geometrical optics and asymptotic matching. Preliminary results show that Theorems 2.2–2.4 can be obtained, though less rigorously, directly from the PDE problem (1)–(3). With the direct methods we are also currently investigating more complex geometries such as ellipses and general convex obstacles.

REFERENCES

[1] J.R. Philip, J.H. Knight, and R.T. Waechter, *Unsaturated seepage and subterranean holes: Conspectus, and exclusion problem for circular cylindrical cavities*, Water Resour. Res., 25 (1989), pp. 16–28.

[2] C. Knessl, *On two-dimensional convection-diffusion past a circle*, SIAM J. Appl. Math., 62 (2001), pp. 310–335.

[3] S.J. Chapman, J.M.H. Lawry, and J.R. Ockendon, *Ray theory for high-Péclet-number convection-diffusion*, SIAM J. Appl. Math., 60 (1999), pp. 121–135.

[4] G.P. Cherepanov, *Two-dimensional convective heat/mass transfer for low Prandtl and any Péclet numbers*, SIAM J. Appl. Math., 58 (1998), pp. 942–960.

[5] J.R. Philip, *Theory of infiltration*, Adv. Hydrosci., 5 (1969), pp. 215–296.

[6] J.H. Knight and J.R. Philip, *The seepage exclusion problem for spherical cavities*, Water Resour. Res., 25 (1989), pp. 29–37.

[7] J.R. Philip, *The seepage exclusion problem for sloping cylindrical cavities*, Water Resour. Res., 25 (1989), pp. 1447–1448.

[8] J.R. Philip, *Asymptotic solutions of the seepage exclusion problem for elliptic-cylindrical, spheroidal, and strip- and disc-shaped cavities*, Water Resour. Res., 25 (1989), pp. 1531–1540.

[9] J.R. Philip, *The seepage exclusion problem for parabolic and paraboidal cavities*, Water Resour. Res., 25 (1989), pp. 605–618.

[10] J.R. Philip, *Some general results on the seepage exclusion problem*, Water Resour. Res., 26 (1990), pp. 369–377.

[11] A. Rosato, F. Prinz, K.J. Standburg, and R. Swendsen, *Monte Carlo simulation of particulate matter segregation*, Powder Tech., 49 (1986), pp. 59–69.

[12] A. Rosato, F. Prinz, K.J. Standburg, and R. Swendsen, *Why the Brazil nuts are on top: Size segregation of particulate matter by shaking*, Phys. Rev. Lett., 58 (1987), pp. 1038–1040.

[13] F.J. Alexander and J.L. Lebowitz, *On the drift and diffusion of a rod in a lattice fluid*, J. Phys. A, 27 (1994), pp. 683–696.

[14] J.J. Bowman, T.B.A. Senior, and P.L.E. Uslenghi, eds., *Electromagnetic and Acoustic Scattering by Simple Shapes*, Hemisphere Publishing, New York, 1987.

[15] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[16] C. Knessl and Y. Yang, *Analysis of a Brownian particle moving in a time-dependent drift field*, Asymptot. Anal., 27 (2001), pp. 281–319.

[17] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed., Academic Press, Boston, 1994.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1580–1597

NEW CONDITIONS ON THE EXISTENCE AND STABILITY OF PERIODIC SOLUTION IN LOTKA–VOLTERRA'S POPULATION SYSTEM[*]

YONGHUI XIA[†] AND MAOAN HAN[‡]

Abstract. In this paper, we revisit the famous periodic Lotka–Volterra competitive system. Some new and interesting sufficient conditions are obtained to guarantee the existence and global asymptotic stability of the periodic solution in the Lotka–Volterra competitive system. Our method is based on Mawhin's coincidence degree, matrix's spectral theory, and some new estimation techniques for the priori bounds of unknown solutions to the equation $Lx = \lambda Nx$. Due to this new method, our new results are much different from the known results in the previous literature. Finally, some examples and their simulations show the feasibility of our results.

Key words. global asymptotic stability, Lotka–Volterra system, periodic solution

AMS subject classifications. 92B20, 34K20, 93D30, 45J05

DOI. 10.1137/070702485

1. Introduction.

In recent years, the application of theories of functional differential equations in mathematical ecology has developed rapidly. Various mathematical models have been proposed in the study of population dynamics, ecology, and epidemiology (see, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]). One of the famous models for dynamics of population is the Lotka–Volterra system. Due to its theoretical and practical significance, the Lotka–Volterra systems have been studied extensively (see, e.g., [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 30, 31, 32, 33, 34]). A basic model is the two-species competitive system which takes the form of

$$\left\{\begin{array}{rcl} \dot{y}_1(t) & = & y_1(t)[b_1 - a_{11}y_1(t) - a_{12}y_2(t)], \\ \dot{y}_2(t) & = & y_2(t)[b_2 - a_{21}y_1(t) - a_{22}y_2(t)]. \end{array}\right. \tag{1}$$

Gopalsamy [8] has studied system (1) and obtained that if $a_{11}b_1 > a_{12}b_2$, $a_{22}b_2 > a_{21}b_1$, $a_{11} > a_{21}$, and $a_{22} > a_{12}$, then system (1) has a unique equilibrium which is globally asymptotically stable.

Then Gopalsamy generalized the results to an n-species competitive system in [9, 10]. The autonomous n-species competitive model can be represented as follows:

$$\dot{y}_i(t) = y_i(t)\left[b_i - \sum_{j=1}^{n} a_{ij}y_j(t)\right], \qquad i = 1, 2, \ldots, n. \tag{2}$$

[*] Received by the editors September 8, 2007; accepted for publication (in revised form) December 15, 2008; published electronically March 11, 2009. This work was supported by Shanghai Leading Academic Discipline Project (S30405), the National Ministry of Education of China (20060270001), the Shanghai Municipal Education Commission under grant (06ZZ16), the National Natural Science Foundation of Shanghai City of China under grant (08ZR1416000), and the Natural Science Foundation of Fujian Province of China under grant (S0750008).

http://www.siam.org/journals/siap/69-6/70248.html

[†] Corresponding Author. Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang, 321004, China (xiadoc@163.com, yhxia02@gmail.com).

[‡] The Institute of Mathematics, Shanghai Normal University, Shanghai, 200234, China (mahan@shnu.edu.cn).

Recently, Xia, Han, and Ding [19] studied systems (1) and (2). By combining matrix's spectral theory with the Lyapunov function, some new sufficient conditions were obtained to guarantee the global asymptotic stability of a unique equilibrium for the Lotka–Volterra system. Their results generalize and significantly improve the known results in [8, 9, 10].

However, most of the literature mentioned above requires the coefficients of the system to be constants. As we know, the variation of the environment plays an important role in many biological and ecological dynamical systems. In particular, the effects of a periodically varying environment are important for evolutionary theory as the selective forces on a system in a fluctuating environment differ from those in a stable environment. If we consider the effects of the environmental factors, the assumption of periodicity of parameters is realistic and important (e.g., seasonal effects of weather, food supplies, mating habits, harvesting, etc.). To incorporate the varying properties of the parameters into the model, many authors considered the nonautonomous n-species competitive system (see, e.g., [2, 3, 4, 5, 7, 9, 11])

$$\dot{y}_i(t) = y_i(t)\left[b_i(t) - \sum_{j=1}^{n} a_{ij}(t)y_j(t)\right], \qquad i = 1, 2, \ldots, n, \tag{3}$$

where the coefficients are assumed to be continuous ω-periodic functions, i.e., $b_i(t+\omega) = b_i(t)$ and $a_{ij}(t+\omega) = a_{ij}(t)$, $i, j = 1, 2, \ldots, n$. For the biological point of view, it is always assumed that the parameters b_i, a_{ij}, $i, j = 1, 2, \ldots, n$, are nonnegative and a_{ii} is strictly positive. System (3) is supplemented with the initial condition

$$y_i(t_0) = y_i^0, \quad y_i^0 > 0, \qquad i = 1, 2, \ldots, n. \tag{4}$$

It is easy to see that for every given positive initial value condition (4), the corresponding solution of (3) remains positive for all $t \geq 0$. That is, the positive cones of R^n are positive invariant with respect to system (3).

To study the existence and global asymptotic stability of periodic solution for system (3), many approaches are employed, such as Brower fixed point, Lyapunov function, comparison theorem, Mawhin's coincidence degree, and so on. In particular, Mawhin's coincidence degree theory is a powerful tool to study the existence of periodic solutions. However, different estimation techniques for the priori bounds of unknown solutions to the equation $Lx = \lambda Nx$ may lead to different results. There are many papers obtaining the priori bounds by employing the inequality

$$|x(t)| \leq |x(t_0)| + \int_0^{\omega} |\dot{x}(t)|dt.$$

For more details, one can refer to [11, 27, 28, 29, 30, 31, 32, 33, 34, 35]. To the best of our knowledge no author has been concerned with employing the matrix's spectral theory to obtain the priori bounds for biological systems so far. For this reason, our aim is to propose a new methodology to revisit system (3). By combing matrix's spectral theory with Mawhin's coincidence degree theory, we manage to obtain a set of new and interesting conditions which are much different from the known results in the literature.

The structure of this paper is as follows. In section 2, some new and interesting sufficient conditions for the existence of periodic solution of system (3) are obtained. Section 3 is devoted to examining the stability of this periodic solution. In section 4, some discussions and remarks on the difference between our results and the previous ones are presented. Finally, some examples and their simulations are given to show the feasibility of our results.

2. Existence of periodic solutions. In this section, we shall obtain some new sufficient conditions for the existence of periodic solution of system (3).

For convenience, we introduce some notation, definitions, and lemmas. If $f(t)$ is a continuous ω-periodic function defined on R, denote

$$\underline{f} = \min_{t\in[0,\omega]} |f(t)|, \qquad \overline{f} = \max_{t\in[0,\omega]} |f(t)|, \qquad m(f) = \frac{1}{\omega}\int_0^{\omega} f(t)dt.$$

For matrix $D = (d_{ij})_{n\times n}$, D^T denotes the transpose of D, and E_n denotes the identity matrix of size n. $\mathrm{diag}(\cdot)$ represents a diagonal matrix with specified diagonal entries. A matrix or vector $A \geq 0$ means that all entries of A are greater than or equal to zero. $A > 0$ can be defined similarly. For matrices or vectors A and B, $A \geq B$ (resp., $A > B$) means that $A - B \geq 0$ (resp., $A - B > 0$). We denote the spectral radius of the matrix A by $\rho(A)$.

Definition 2.1 (see [34, 35]). *Let X, Z be normed real Banach spaces, let $L : \mathrm{Dom}L \subset X \to Z$ be a linear mapping, and let $N : X \to Z$ be a continuous mapping. The mapping L is called a Fredholm mapping of index zero if* $\mathrm{dim}KerL = \mathrm{codim}\ \mathrm{Im}L < +\infty$ *and* $\mathrm{Im}L$ *is closed in Z. If L is a Fredholm mapping of index zero and there exist continuous projectors $P : X \to X$ and $Q : Z \to Z$ such that* $\mathrm{Im}P = \mathrm{Ker}L$, $\mathrm{Ker}Q = \mathrm{Im}L = \mathrm{Im}(I - Q)$, *it follows that* $L|\mathrm{dom}L \cap \mathrm{Ker}P : (I - P)X \to \mathrm{Im}L$ *is invertible. We denote the inverse of that map by K_P. If Ω is an open bounded subset of X, the mapping N will be called L-compact on $\overline{\Omega}$ if $QN(\overline{\Omega})$ is bounded and $K_P(I - Q)N : \overline{\Omega} \to X$ is compact. Since* $\mathrm{Im}Q$ *is isomorphic to* $\mathrm{Ker}L$, *there exists an isomorphism* $J : \mathrm{Im}Q \to \mathrm{Ker}L$.

Definition 2.2 (see [35]). *Let $\Omega \subset R^n$ be open and bounded, $f \in C^1(\Omega, R^n) \cap C(\overline{\Omega}, R^n)$, and $y \in R^n/f(\partial\Omega \cup N_f)$, i.e., y is a regular value of f. Here, $N_f = \{x \in \Omega : J_f(x) = 0\}$, the critical set of f, and J_f is the Jacobian of f at x. Then the degree $deg\{f, \Omega, y\}$ is defined by*

$$\deg\{f, \Omega, y\} = \sum_{x\in f^{-1}(y)} \mathrm{sgn} J_f(x)$$

with the agreement that $\sum \phi = 0$. For more details about degree theory, the reader is referred to [35].

Lemma 2.1 (continuation theorem [34]). *Let $\Omega \subset X$ be an open and bounded set. Let L be a Fredholm mapping of index zero, and let N be L-compact on $\overline{\Omega}$ (i.e., $QN(\overline{\Omega})$ is bounded and $K_P(I - Q)N : \overline{\Omega} \to X$ is compact). Assume the following:*

(i) *For each $\lambda \in (0, 1)$, $x \in \partial\Omega \cap \mathrm{Dom}L$, $Lx \neq \lambda Nx$.*

(ii) *For each $x \in \partial\Omega \cap \mathrm{Ker}L$, $QNx \neq 0$ and* $\deg\{JQN,\ \Omega \cap \mathrm{Ker}L, 0\} \neq 0$.

Then $Lx = Nx$ has at least one solution in $\overline{\Omega} \cap \mathrm{Dom}L$.

Definition 2.3 (see [25, 42]). *A real $n \times n$ matrix $A = (a_{ij})$ is said to be an M-matrix if $a_{ij} \leq 0$, $i, j = 1, 2, \ldots, n$, $i \neq j$, and $A^{-1} \geq 0$.*

Lemma 2.2 (see [25, 26]). *Let $A \geq 0$ be an $n \times n$ matrix and let $\rho(A) < 1$. Then $(E_n - A)^{-1} \geq 0$, where E_n denotes the identity matrix of size n.*

Lemma 2.3 (see [20, 29]). *Assume that*

$$m(b_i) > \sum_{j=1,j\neq i}^{n} m(a_{ij}) \frac{m(b_j)}{m(a_{jj})}, \qquad i = 1, 2, \ldots, n.$$

Then the system of algebraic equations

$$\sum_{j=1}^{n} m(a_{ij})u_j = m(b_i), \qquad i = 1, 2, \dots, n,$$

has a unique solution $(u_1^*, u_2^*, \dots, u_n^*)^T \in R_+^n$ *with* $u_i^* > 0$.

In what follows, we shall introduce some function spaces and their norms, which are valid throughout this paper. Denote

$$X = \{x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T \in C^1(R, R^n) | x(t+\omega) = x(t) \text{ for all } t \in R\},$$

$$Z = \{x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T \in C(R, R^n) | x(t+\omega) = x(t) \text{ for all } t \in R\}.$$

The norms are given by

$$|x_i(t)|_0 = \max_{t\in[0,\omega]} |x_i(t)|, \qquad |x_i(t)|_1 = |x_i(t)|_0 + |\dot{x}_i(t)|_0, \quad i = 1, 2, \dots, n,$$

$$\|x(t)\|_0 = \max_{1\le i\le n} \{|x_i(t)|_0\}, \qquad \|x(t)\|_1 = \|x(t)\|_0 + \|\dot{x}(t)\|_0 = \max_{1\le i\le n} \{|x_i(t)|_1\}.$$

Obviously, X and Z, respectively endowed with the norms $\|\cdot\|_1$ and $\|\cdot\|_0$, are Banach spaces.

THEOREM 2.1. *Assume that the following conditions hold:*

(H₁) $m(b_i) > \sum_{j=1, j\neq i}^{n} m(a_{ij}) \dfrac{m(b_j)}{m(a_{jj})}$, $i = 1, 2, \dots, n$;

(H₂) $\rho(\mathcal{K}) < 1$, *where* $\mathcal{K} = (\Gamma_{ij})_{n\times n}$ *and* $\Gamma_{ij} = \begin{cases} 0, & i = j, \\ \overline{a}_{ij}\underline{a}_{jj}^{-1}, & i \neq j. \end{cases}$

Then system (3) *has at least one positive* ω*-periodic solution.*

Proof. Note that every solution $y(t) = (y_1(t), y_2(t), \dots, y_n(t))^T$ of system (3) with the initial value condition (4) is positive. Make the change of variables

$$x_i(t) = \ln y_i(t), \quad i = 1, 2, \dots, n. \tag{5}$$

Then system (3) can be rewritten as

$$\dot{x}_i(t) = b_i(t) - \sum_{j=1}^{n} a_{ij}(t) e^{x_j(t)}, \quad i = 1, 2, \dots, n. \tag{6}$$

Obviously, system (3) has at least one ω-periodic solution which is equivalent so that system (6) has at least one ω-periodic solution. To prove Theorem 2.1, our main tasks are to construct the operators (i.e., L, N, P, and Q) appearing in Lemma 2.1 and to find an appropriate open set Ω satisfying conditions (i), (ii) in Lemma 2.1. To this end, we proceed with three steps.

Step 1. In this step, we intend to construct the operators appearing in Lemma 2.1 and verify that they satisfy the conditions of Lemma 2.1. For any $x(t) \in X$, in view of the periodicity, it is easy to check that

$$\Delta_i(x, t) = b_i(t) - \sum_{j=1}^{n} a_{ij}(t) e^{x_j(t)} \in Z.$$

Now we define the operators $L : \mathrm{DomL} \subset \mathrm{X} \to \mathrm{Z}$ and $N : X \to Z$ as follows:

$$X \ni x(t) \to (Lx)(t) = \frac{\mathrm{d}x(t)}{\mathrm{d}t} \in Z,$$

$$X \ni x(t) \to (Nx)(t) = \big((Nx)_1(t), (Nx)_2(t), \ldots, (Nx)_n(t)\big)^T \in Z,$$

where

$$(Nx)_i(t) = \Delta_i(x,t), \quad i = 1, 2, \ldots, n.$$

Define, respectively, the projectors $P : X \to X$ and $Q : Z \to Z$ by

$$Px(t) = \frac{1}{\omega}\int_0^\omega x(t)dt, \qquad Qz(t) = \frac{1}{\omega}\int_0^\omega z(t)dt, \quad x \in X,\ z \in Z.$$

It is obvious that the domain of L in X is actually the whole space, and

$$\mathrm{Ker}L = \{x(t) \in X | Lx(t) = 0,\ \text{i.e., } \dot{x}(t) = 0\} = R^n,$$

$$\mathrm{Im}L = \left\{z(t) \in Z | \int_0^\omega z(t)dt = 0\right\} \text{ is closed in } Z.$$

Moreover, P, Q are continuous operators such that

$$\mathrm{Im}P = R^n = \mathrm{Ker}L, \qquad \mathrm{Im}L = \mathrm{Ker}Q = \mathrm{Im}(I - Q),$$

and

$$\mathrm{dimKer}L = \mathrm{codimIm}L = n < +\infty.$$

It follows that L is a Fredholm mapping of index zero. Furthermore, the generalized inverse (to L) $K_P : \mathrm{Im}L \to \mathrm{Dom}L \cap \mathrm{Ker}P$ exists, which is given by

$$K_P(y) = \int_0^t y(s)ds - \frac{1}{\omega}\int_0^\omega \int_0^t y(s)dsdt.$$

Then $QN : X \to Z$ and $K_P(I - Q)N : X \to X$ are defined by

$$QNx = \left(\frac{1}{\omega}\int_0^\omega \Delta_1(x,t)\mathrm{d}t, \frac{1}{\omega}\int_0^\omega \Delta_2(x,t)\mathrm{d}t, \ldots, \frac{1}{\omega}\int_0^\omega \Delta_n(x,t)\mathrm{d}t\right)^T,$$

$$K_P(I - Q)Nx = (\Psi_1(x,t), \Psi_2(x,t), \ldots, \Psi_n(x,t))^T, \tag{7}$$

where

$$\Psi_k(x,t) = \int_0^t \Delta_k(x,u)du - \frac{1}{\omega}\int_0^\omega \int_0^t \Delta_k(x,u)dudt - \Big(\frac{t}{\omega} - \frac{1}{2}\Big)\int_0^\omega \Delta_k(x,u)du,$$
$$k = 1, 2, \ldots, n.$$

Clearly, QN and $K_P(I - Q)N$ are continuous. Now we turn to the fact that for any open bounded set $\Omega \subset X$, denoted by

$$\Omega = \{x(t) \in X \big| |x_i(t)|_1 = |x_i(t)|_0 + |\dot{x}_i(t)|_0 < h_i\},$$

the mapping N is L-compact on $\overline{\Omega}$. Here, the constants h_i are independent of the choice of $x(t)$. In view of Definition 2.1, to show the above fact, it suffices to show that $QN(\overline{\Omega})$ is bounded and $K_P(I-Q)N : \overline{\Omega} \to X$ is compact. We first arrive at

$$|(QNx)_i|_0 = \left|\frac{1}{\omega}\int_0^\omega \Delta_i(x,t)dt\right|_0 \leq |\Delta_i(x,t)|_0 := M_i \text{ for all } x \in \overline{\Omega},$$

which implies that $QN(\overline{\Omega})$ is bounded in the space $(Z, \|\cdot\|_0)$. Second, we shall show that $\big(K_P(I-Q)Nx\big)(\overline{\Omega})$ is relatively compact in the space $(X, \|\cdot\|_1)$. In fact, it follows from (7) that

$$(K_P(I-Q)Nx)' = (\Psi_1'(x,t), \Psi_2'(x,t), \ldots, \Psi_n'(x,t))^T, \tag{8}$$

where $' = d/dt$ and

$$\Psi_k'(x,t) = \Delta_k(x,t) - \frac{1}{\omega}\int_0^\omega \Delta_k(x,u)du, \quad k = 1, 2, \ldots, n.$$

This, combined with (7), gives

$$\begin{aligned} |\big(K_P(I-Q)Nx\big)_i(t)|_1 &= |\big(K_P(I-Q)Nx\big)_i(t)|_0 + |\big(K_P(I-Q)Nx\big)_i'(t)|_0 \\ &\leq M_i\omega + \tfrac{1}{2}M_i\omega + \tfrac{1}{2}M_i\omega + M_i + M_i = 2(\omega+1)M_i, \end{aligned}$$

which implies that $K_P(I-Q)N(\overline{\Omega})$ is bound in the space $(X, \|\cdot\|_1)$.

On the other hand, we prove that $\big(K_P(I-Q)Nx\big)(\overline{\Omega})$ is equicontinuous. In view of the uniform continuity of b_i and a_{ij}, for any $\varepsilon > 0$, there exists $\delta_1 > 0$ such that for any $t, s \in R$, provided that $|t-s| < \delta_1$, we have

$$|b_i(t) - b_i(s)| < \varepsilon, \quad |a_{ij}(t) - a_{ij}(s)| < \varepsilon. \tag{9}$$

Since any $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T \in \overline{\Omega}$ is equicontinuous, for the same ε, there exists $0 < \delta_2 \leq \delta_1$ such that for any $t, s \in R$, provided that $|t-s| < \delta_2$, we have

$$|x_i(t) - x_i(s)|_0 < \varepsilon. \tag{10}$$

It follows from (9) and (10) that

$$\begin{aligned} & |\Delta_i(x(t),t) - \Delta_i(x(s),s)|_0 \\ \leq\ & |b_i(t) - b_i(s)|_0 + \sum_{j=1}^n |a_{ij}(t)e^{x_j(t)} - a_{ij}(s)e^{x_j(s)}|_0 \\ \leq\ & |b_i(t) - b_i(s)| + \sum_{j=1}^n |a_{ij}(t)|_0 |e^{x_j(t)} - e^{x_j(s)}|_0 + \sum_{j=1}^n |a_{ij}(t) - a_{ij}(s)|_0 |e^{x_j(s)}|_0 \\ <\ & \varepsilon + \sum_{j=1}^n \overline{a}_{ij} e^{h_j} |x_i(t) - x_i(s)|_0 + \sum_{j=1}^n e^{h_j}\varepsilon \\ <\ & \varepsilon + \sum_{j=1}^n \overline{a}_{ij} e^{h_j}\varepsilon + \sum_{j=1}^n e^{h_j}\varepsilon \\ =\ & \left[1 + \sum_{j=1}^n (1 + \overline{a}_{ij})e^{h_j}\right]\varepsilon. \end{aligned}$$

Thus, it follows from (8) that

$$\begin{aligned} & |\big(K_P(I-Q)Nx\big)_i'(t) - \big(K_P(I-Q)Nx\big)_i'(s)|_0 \\ =\ & |\Delta_i(x(t),t) - \Delta_i(x(s),s)|_0 \\ <\ & \left[1 + \sum_{j=1}^n (1 + \overline{a}_{ij})e^{h_j}\right]\varepsilon. \end{aligned} \tag{11}$$

On the other hand, the mean value theorem together with (8) gives

$$\begin{aligned}&|\big(K_P(I-Q)Nx\big)_i(t)-\big(K_P(I-Q)Nx\big)_i(s)|_0\\ =\;&\|\big(K_P(I-Q)Nx\big)_i'(\xi)\|_0|t-s|\leq 2M_i|t-s|,\end{aligned}\tag{12}$$

where ξ lies between t and s. Taking $\delta=\min\{\frac{\varepsilon}{2M_i},\delta_2\}$, it follows from (11) and (12) that $|t-s|<\delta$ implies

$$\begin{aligned}&|\big(K_P(I-Q)Nx\big)_i(t)-\big(K_P(I-Q)Nx\big)_i(s)|_1\\ =\;&|\big(K_P(I-Q)Nx\big)_i(t)-\big(K_P(I-Q)Nx\big)_i(s)|_0\\ &+|\big(K_P(I-Q)Nx\big)_i'(t)-\big(K_P(I-Q)Nx\big)_i'(s)|_0\\ <\;&\left[1+\sum_{j=1}^{n}(1+\overline{a}_{ij})e^{h_j}\right]\varepsilon+2M_i\delta\\ <\;&\left[1+\sum_{j=1}^{n}(1+\overline{a}_{ij})e^{h_j}\right]\varepsilon+\varepsilon:=\widetilde{M}\varepsilon,\end{aligned}$$

which implies that $\big(K_P(I-Q)Nx\big)(\overline{\Omega})$ is equicontinuous.

Therefore, by the generalized Arzela–Ascoli theorem, we have that $\big(K_P(I-Q)Nx\big)(\overline{\Omega})$ is relatively compact in the space $(X,\|\cdot\|_1)$. The proof of this step is complete.

Step 2. In this step, we are in a position to search for an appropriate open bounded subset Ω satisfying condition (i) of Lemma 2.1. Specifically, our aim is to search for an appropriate h_i defined by Ω in Step 1 such that Ω satisfies condition (i) of Lemma 2.1. To this end, assume that $x(t)\in X$ is a solution of the equation $Lx=\lambda Nx$ for each $\lambda\in(0,1)$; that is,

$$\dot{x}_i(t)=\lambda\left[b_i(t)-\sum_{j=1}^{n}a_{ij}(t)e^{x_j(t)}\right],\quad i=1,2,\dots,n.\tag{13}$$

Since $x(t)\in X$, each $x_i(t)$, $i=1,2,\dots,n$, as components of $x(t)$, is continuously differentiable and ω-periodic. In view of continuity and periodicity, there exists $t_i\in[0,\omega]$ such that $x_i(t_i)=\max_{t\in[0,\omega]}|x_i(t)|$, $i=1,2,\dots,n$. Accordingly, $\dot{x}_i(t_i)=0$, and we arrive at

$$b_i(t_i)-\sum_{j=1}^{n}a_{ij}(t_i)e^{x_j(t_i)}=0,\quad i=1,2,\dots,n.$$

That is,

$$a_{ii}(t_i)e^{x_i(t_i)}=b_i(t_i)-\sum_{j=1,j\neq i}^{n}a_{ij}(t_i)e^{x_j(t_i)},\quad i=1,2,\dots,n.$$

Noticing that $x_j(t_j)=\max_{t\in[0,\omega]}|x_j(t)|$ implies $x_j(t_i)\leq x_j(t_j)$, it follows that

$$\begin{aligned}\underline{a}_{ii}e^{x_i(t_i)}\leq\left|a_{ii}(t_i)e^{x_i(t_i)}\right|\;&=\;\left|b_i(t_i)-\sum_{j=1,j\neq i}^{n}a_{ij}(t_i)e^{x_j(t_i)}\right|\\ &\leq\;\overline{b}_i+\sum_{j=1,j\neq i}^{n}\overline{a}_{ij}e^{x_j(t_i)}\leq\overline{b}_i+\sum_{j=1,j\neq i}^{n}\overline{a}_{ij}e^{x_j(t_j)}.\end{aligned}\tag{14}$$

Letting $\underline{a}_{ii}e^{x_i(t_i)} = z_i(t_i)$, it follows from (14) that

$$z_i(t_i) \le \overline{b}_i + \sum_{j=1, j\neq i}^{n} \overline{a}_{ij}\underline{a}_{jj}^{-1} z_j(t_j)$$

or

$$z_i(t_i) - \sum_{j=1, j\neq i}^{n} \overline{a}_{ij}\underline{a}_{jj}^{-1} z_j(t_j) \le \overline{b}_i,$$

which implies

$$\left(\begin{array}{llll} \;\;1 & -\underline{a}_{22}^{-1}\overline{a}_{12} & \dots & -\underline{a}_{nn}^{-1}\overline{a}_{1n} \\ -\underline{a}_{11}^{-1}\overline{a}_{21} & \;\;1 & \dots & -\underline{a}_{nn}^{-1}\overline{a}_{2n} \\ \dots & \dots & \dots & \dots \\ -\underline{a}_{11}^{-1}\overline{a}_{n1} & -\underline{a}_{22}^{-1}\overline{a}_{n2} & \dots & \;\;1 \end{array}\right) \left(\begin{array}{c} z_1(t_1) \\ z_2(t_2) \\ \dots \\ z_n(t_n) \end{array}\right) \le \left(\begin{array}{c} \overline{b}_1 \\ \overline{b}_2 \\ \dots \\ \overline{b}_n \end{array}\right). \tag{15}$$

Set $D = (D_1, D_2, \dots, D_n)^T = (\overline{b}_1, \overline{b}_2, \dots, \overline{b}_n)^T$. It follows from (15) that

$$(E - \mathcal{K})\big(z_1(t_1), z_2(t_2), \dots, z_n(t_n)\big)^T \le D. \tag{16}$$

In view of $\rho(\mathcal{K}) < 1$ and Lemma 2.2, $(E_n - \mathcal{K})^{-1} \ge 0$. Let

$$H = (\widetilde{h}_1, \widetilde{h}_2, \dots, \widetilde{h}_n)^T := (E - \mathcal{K})^{-1} D \ge 0. \tag{17}$$

Then it follows from (16) and (17) that

$$\big(z_1(t_1), z_2(t_2), \dots, z_n(t_n)\big)^T \le H, \text{ or } z_i(t_i) \le \widetilde{h}_i, \quad i = 1, 2, \dots, n, \tag{18}$$

which implies

$$|x_i(t)|_0 = \max_{t\in[0,\omega]} |x_i(t)| = x_i(t_i) \le \ln \frac{\widetilde{h}_i}{\underline{a}_{ii}}, \quad i = 1, 2, \dots, n.$$

On the other hand, it follows from (17) that

$$(E - \mathcal{K})H = D, \text{ or } H = \mathcal{K}H + D, \text{ that is,} \tag{19}$$
$$\widetilde{h}_i = \sum_{j=1}^{n} \Gamma_{ij}\widetilde{h}_j + D_i, \; i = 1, 2, \dots, n.$$

Estimating (13), by using (18) and (19), we have

$$\begin{array}{rcl} |\dot{x}_i(t)|_0 & = & \lambda\Big|b_i(t) - \sum\limits_{j=1}^{n} a_{ij}(t)e^{x_j(t)}\Big|_0 \\ & \le & \overline{b}_i + \sum\limits_{j=1}^{n} \overline{a}_{ij}|e^{x_j(t)}|_0 = \overline{b}_i + \sum\limits_{j=1}^{n} \overline{a}_{ij}e^{x_j(t_j)} \\ & \le & \overline{b}_i + \sum\limits_{j=1}^{n} \overline{a}_{ij}\underline{a}_{jj}^{-1} z_j(t_j) \\ & = & \overline{b}_i + \sum\limits_{j=1, j\neq i}^{n} \overline{a}_{ij}\underline{a}_{jj}^{-1} z_j(t_j) + z_i(t_i) \\ & \le & D_i + \sum\limits_{j=1}^{n} \Gamma_{ij}\widetilde{h}_j + \widetilde{h}_i \\ & \le & \widetilde{h}_i + \widetilde{h}_i = 2\widetilde{h}_i. \end{array} \tag{20}$$

We can choose a large enough real number ($d > 1$) such that

$$\ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}} > \ln \frac{\widetilde{h}_i}{\underline{a}_{ii}} + 2\widetilde{h}_i.$$

Set $h_i = \ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}}$. Then for any solution of $Lx = \lambda Nx$, we have $|x_i(t)|_1 = |x_i(t)|_0 + |\dot{x}_i(t)|_0 \leq \ln \frac{\widetilde{h}_i}{\underline{a}_{ii}} + 2\widetilde{h}_i < h_i$ for all $i = 1, 2, \ldots, n$. Obviously, h_i are independent of λ and the choice of $x(t)$. Consequently, taking $h_i = \ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}}$, the open subset Ω satisfies that $Lx \neq \lambda Nx$ for each $\lambda \in (0, 1)$, $x \in \partial\Omega \cap \mathrm{Dom}L$, i.e., the open subset Ω satisfies the assumption (i) of Lemma 2.1.

Step 3. In what follows, we verify that for the given open bounded set Ω obtained in Step 2, the assumption (ii) of Lemma 2.1 also holds. That is, for each $x \in \partial\Omega \cap \mathrm{Ker}L$, $QNx \neq 0$, and $\deg\{JQN,\ \Omega \cap \mathrm{Ker}L, 0\} \neq 0$.

Take $x \in \partial\Omega \cap \mathrm{Ker}L$. Then, in view of $\mathrm{Ker}L = R^n$, x is a constant vector in R^n, denoted by $x = (x_1, x_2, \ldots, x_n)^T$ and with the property

$$|x_i| = |x_i|_0 = |x_i|_1 = h_i = \ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}} > \ln \frac{\widetilde{h}_i}{\underline{a}_{ii}} + 2\widetilde{h}_i \text{ for all } i = 1, 2, \ldots, n. \tag{21}$$

Operate x by QN, and we obtain that for $i = 1, 2, \ldots, n$,

$$(QNx)_i = m(b_i) - \sum_{j=1}^{n} m(a_{ij}) e^{x_j}, \quad i = 1, 2, \ldots, n.$$

We claim that $|(QNx)_i| > 0$ for $i = 1, 2, \ldots, n$. If this is not valid, suppose that there exists a certain $k \in \{1, 2, \ldots, n\}$ such that $|(QNx)_k| = 0$, i.e.,

$$m(b_k) - \sum_{j=1}^{n} m(a_{kj}) e^{x_j} = 0 \text{ or } m(a_{kk}) e^{x_k} = m(b_k) - \sum_{j=1, j\neq k}^{n} m(a_{kj}) e^{x_j}.$$

Letting $m(a_{kk}) e^{x_k} = y_k$, we have

$$y_k = m(b_k) - \sum_{j=1, j\neq k}^{n} \frac{m(a_{kj})}{m(a_{jj})} y_j. \tag{22}$$

In view of (21), we get
(23)

$$|y_i| = |y_i|_0 = |y_i|_1 = m(a_{ii}) e^{h_i} = m(a_{ii}) e^{\ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}}} = m(a_{ii}) \frac{d\widetilde{h}_i}{\underline{a}_{ii}} \text{ for all } i = 1, 2, \ldots, n.$$

Note that $\underline{f} \leq m(f) \leq \overline{f}$. It follows from (22), (23), and (19) that

$$\begin{aligned}
d\widetilde{h}_k \leq m(a_{kk})\frac{d\widetilde{h}_k}{\underline{a}_{kk}} = |y_k| &= \left| m(b_k) - \sum_{j=1,j\neq k}^{n} \frac{m(a_{kj})}{m(a_{jj})} y_j \right| \\
&\leq \sum_{j=1,j\neq k}^{n} \frac{m(a_{kj})}{m(a_{jj})} |y_j| + \overline{b}_k \\
&\leq \sum_{j=1,j\neq k}^{n} \frac{m(a_{kj})}{m(a_{jj})} m(a_{jj}) \frac{d\widetilde{h}_j}{\underline{a}_{jj}} + \overline{b}_k \\
&\leq \sum_{j=1,j\neq k}^{n} \overline{a}_{kj} \underline{a}_{jj}^{-1} d\widetilde{h}_j + D_k \\
&< d \sum_{j=1,j\neq k}^{n} \overline{a}_{kj} \underline{a}_{jj}^{-1} \widetilde{h}_j + dD_k \\
&< d \Bigg[\sum_{j=1,j\neq k}^{n} \Gamma_{kj} \widetilde{h}_j + D_k \Bigg] \\
&= d\widetilde{h}_k,
\end{aligned}$$

which is a contradiction. Therefore, for any $x \in \partial\Omega \cap \mathrm{Ker}L$, $|(QNx)_i| > 0$ for all $i = 1, 2, \ldots, n$. That is, $(QNx) \neq 0$ for $x \in \partial\Omega \cap \mathrm{Ker}L$.

Next, we show that the topological degree is nonzero. In fact, it follows from (H$_1$) and Lemma 2.3 that the algebraic equation

$$\sum_{j=1}^{n} m(a_{ij}) u_j = m(b_i), \quad i = 1, 2, \ldots, n,$$

has a unique solution $u^* = (u_1^*, u_2^*, \ldots, u_n^*)^T \in R_+^n$ with $u_i^* > 0$. Obviously, the algebraic equation

$$\sum_{j=1}^{n} m(a_{ij}) e^{v_j} = m(b_i), \quad i = 1, 2, \ldots, n, \tag{24}$$

has a unique solution $v^* = (v_1^*, v_2^*, \ldots, v_n^*)^T \in R^n$. Now we further claim that the unique solution $v^* \in \Omega \cap \mathrm{Ker}L$. Indeed, if this is not the case, suppose that

$$|v_i^*| = |v_i^*|_0 = |v_i^*|_1 = h_i = \ln \frac{d\widetilde{h}_i}{\underline{a}_{ii}} > \ln \frac{\widetilde{h}_i}{\underline{a}_{ii}} + 2\widetilde{h}_i \text{ for all } i = 1, 2, \ldots, n. \tag{25}$$

Letting $m(a_{ii})e^{v_i} = u_i$, we have

$$u_i = m(b_i) - \sum_{j=1,j\neq i}^{n} \frac{m(a_{ij})}{m(a_{jj})} u_j. \tag{26}$$

By using (25), (26), and (19), a similar argument in the above leads to

$$\begin{aligned}
d\widetilde{h}_i \leq m(a_{ii})\frac{d\widetilde{h}_i}{\underline{a}_{ii}} = |u_i| &= \left| m(b_i) - \sum_{j=1,j\neq i}^{n} \frac{m(a_{ij})}{m(a_{jj})} u_j \right| \\
&< d \left[\sum_{j=1,j\neq k}^{n} \Gamma_{ij} \widetilde{h}_j + D_i \right] = d\widetilde{h}_i,
\end{aligned}$$

which is a contradiction. Therefore, for any $v^* \in \Omega \cap \mathrm{Ker}L$, in view of Definition 2.2, it is easy to see that

$$\deg\{\,\mathrm{JQN}, \Omega \cap \mathrm{Ker}L, 0\} = \mathrm{sign}\left[(-1)^n \det[m(a_{ij})] \exp\left\{\sum_{i=1}^{n} v_j^*\right\}\right] \neq 0,$$

where $\deg(\cdot)$ is the Brouwer degree and J is the identity mapping since $\mathrm{Im}Q = \mathrm{Ker}L$.

We have shown that the open subset $\Omega \subset X$ satisfies all the assumptions of Lemma 2.1. Hence, by Lemma 2.1, system (6) has at least one positive ω-periodic solution in $\mathrm{Dom}L \cap \overline{\Omega}$. By (5), system (3) has at least one positive ω-periodic solution. This completes the proof of Theorem 2.1. □

3. Globally asymptotic stability. Under the assumption of Theorem 2.1, we know that system (3) has at least one positive ω-periodic solution, denoted by $y^*(t) = \left(y_1^*(t), \ldots, y_n^*(t)\right)^T$. The aim of this section is to derive a set of sufficient conditions which guarantee the existence and global asymptotic stability of the positive ω-periodic solution $y^*(t)$.

Before the formal analysis, we recall some facts which will be used in the proof.

DEFINITION 3.1. *Let* $y^*(t) = \left(y_1^*(t), \ldots, y_n^*(t)\right)^T$ *be the* ω*-periodic solution of* (3) *and let* $y(t) = \left(y_1(t), \ldots, y_n(t)\right)^T$ *be any positive solution of* (3). *We say* $y^*(t)$ *is globally asymptotically stable if the following conditions hold:*

(i) $y^*(t)$ *is Lyapunov stable;*

(ii) $y^*(t)$ *is globally attractive in the sense that* $\lim_{t\to+\infty}[y_i(t) - y_i^*(t)] = 0$ *for all* $i = 1, 2, \ldots, n$.

LEMMA 3.1 (see [22, 24, 30]). *Let f be a nonnegative function defined on* $[0, +\infty]$ *such that f is integrable on* $[0, +\infty]$ *and is uniformly continuous on* $[0, +\infty]$. *Then* $\lim_{t\to+\infty} f(t) = 0$.

LEMMA 3.2 (see [25, 26, 42, 43]). *Let* $\mathcal{K} = (\Gamma_{ij})_{n\times n}$ *be a matrix with nonpositive off-diagonal elements.* $\mathcal{K}$ *is an* M*-matrix if and only if there exists a positive diagonal matrix* $\xi = \mathrm{diag}(\xi_1, \xi_2, \ldots, \xi_n)$ *such that*

$$\xi_i \underline{a}_{ii} > \sum_{j\neq i} \xi_j \overline{a}_{ij}, \quad i = 1, 2, \ldots, n.$$

THEOREM 3.1. *Assume that all the assumptions in Theorem* 2.1 *hold. Then system* (3) *has a unique positive* ω*-periodic solution* $y^*(t)$ *which is globally asymptotically stable.*

Proof. Let $y(t) = \left(y_1(t), \ldots, y_n(t)\right)^T$ be any positive solution of system (3). Set

$$Y_i(t) = \ln y_i(t), \quad Y_i^*(t) = \ln y_i^*(t). \tag{27}$$

Then, it follows from (27) and (3) that

$$D^+\left[|Y_i(t) - Y_i^*(t)|\right] \leq -a_{ii}(t)|y_i(t) - y_i^*(t)| + \sum_{j=1, j\neq i}^{n} a_{ij}(t)|y_j(t) - y_j^*(t)|. \tag{28}$$

It is easy to see that $\rho(\mathcal{K}^T) = \rho(\mathcal{K}) < 1$. Thus, in view of Lemma 2.2 and Definition 2.3, $(E - \mathcal{K}^T)$ is an M-matrix, where E denotes an identity matrix of size n. Therefore, by Lemma 3.2, there exists a diagonal matrix $\xi = \mathrm{diag}(\xi_1, \ldots, \xi_n)$ with positive diagonal elements such that the product $(E - \mathcal{K}^T)\xi$ is strictly diagonally dominant with positive diagonal entries, namely,

$$\xi_i \Gamma_{ii} > \sum_{j=1, j\neq i}^{n} \xi_j \Gamma_{ji} \text{ or } \xi_i \underline{a}_{ii} - \sum_{j=1, j\neq i}^{n} \xi_j \overline{a}_{ji} > 0, \quad i = 1, \dots, n. \tag{29}$$

Now, we define a Lyapunov function $V(t)$ as follows:

$$V(t) = \sum_{i=1}^{n} \xi_i |Y_i(t) - Y_i^*(t)|, \quad t \geq t_0. \tag{30}$$

Let $Z_i(t) = |y_i(t) - y_i^*(t)|$. Calculating the upper right derivative of $V(t)$, it follows from (30) and (28) that

$$\begin{aligned} D^+V(t) &\leq \sum_{i=1}^{n} \xi_i \left\{ -a_{ii}(t)|y_i(t) - y_i^*(t)| + \sum_{j\neq i, j=1}^{n} a_{ij}(t)|y_j(t) - y_j^*(t)| \right\} \\ &= -\sum_{i=1}^{n} \xi_i a_{ii}(t) Z_i(t) + \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} \xi_j a_{ji}(t) Z_i(t) \\ &= -\sum_{i=1}^{n} \left\{ \xi_i a_{ii}(t) - \sum_{j\neq i j=1}^{n} \xi_j a_{ji}(t) \right\} Z_i(t) \\ &\leq -c \sum_{i=1}^{n} Z_i(t) \leq 0, \quad t \geq t_0, \end{aligned} \tag{31}$$

where $c = \max_{1\leqslant i\leqslant n} \sup_{t\in[0,\omega]} \{\xi_i a_{ii}(t) - \sum_{j=1, j\neq i}^{n} \xi_j a_{ji}(t)\} > 0$. It follows from (31) that $D^+V(t) \leq 0$. Obviously, the zero solution of (3) is Lyapunov stable. On the other hand, integrating (31) over $[t_0, t]$ leads to

$$V(t) - V(t_0) \leqslant -c \int_{t_0}^{t} \sum_{i=1}^{n} Z_i(s) ds, \qquad t \geqslant 0,$$

or

$$V(t) + c \int_{t_0}^{t} \sum_{i=1}^{n} |y_i(s) - y_i^*(s)| ds \leqslant V(t_0) < +\infty, \quad t \geqslant t_0.$$

Noting that $V(t) \geqslant 0$, it follows that

$$\int_{t_0}^{t} \sum_{i=1}^{n} |y_i(s) - y_i^*(s)| ds \leqslant \frac{V(t_0)}{c} < +\infty, \quad t \geqslant t_0. \tag{32}$$

Therefore, by Lemma 3.1, it is not difficult to conclude that

$$\lim_{t\to+\infty} |y_i(t) - y_i^*(t)| = 0.$$

From Definition 3.1, Theorem 3.1 follows. □

4. Corollaries, remarks, and conclusions. In order to illustrate some features of our main results, we will present some corollaries and remarks in this section.

From the proofs of Theorems 2.1 and 3.1, a direct corollary follows immediately.

Corollary 4.1. *In addition to* (H_1)*, suppose further that* $E-\mathcal{K}$ *or* $E-\mathcal{K}^T$ *is an* M*-matrix. Then system* (3) *has a unique positive* ω*-periodic solution which is globally asymptotically stable.*

Now recall that for a given matrix $\mathcal{K}$, its spectral radius $\rho(\mathcal{K})$ is equal to the minimum of all matrix norms of $\mathcal{K}$, i.e., for any matrix norm $\|\cdot\|$, $\rho(\mathcal{K}) \le \|\mathcal{K}\|$. Therefore, we have the following corollary.

Corollary 4.2. *In addition to* (H_1)*, suppose further that there exist positive constants* ξ_i, $i=1,2,\ldots,n$ *such that one of the following conditions holds:*

(1) $\max\limits_{1\le j\le n}[\underline{a}_{jj}^{-1}\xi_j^{-1}\sum_{i=1,i\neq j}^{n}\xi_i\overline{a}_{ij}]<1$*, or, equivalently,* $\underline{a}_{jj}>\sum_{i=1,i\neq j}^{n}\overline{a}_{ij}$ *for all* $j=1,2,\ldots,n$.

(2) $\sum_{i=1}^{n}\sum_{j=1}^{n}(\xi_i^{-1}\xi_j\Gamma_{ij})^2<1$*, where*

$$\Gamma_{ij}=\begin{cases}0, & i = j,\\ \underline{a}_{jj}^{-1}\overline{a}_{ij}, & i \neq j.\end{cases}$$

(3) $\max\limits_{1\le i\le n}[\underline{a}_{ii}^{-1}\xi_i^{-1}\sum_{j=1,j\neq i}^{n}\xi_j\overline{a}_{ji}]<1$*, or, equivalently,* $\underline{a}_{ii}>\sum_{j=1,j\neq i}^{n}\overline{a}_{ji}$ *for all* $i=1,2,\ldots,n$.

Then system (3) *has a unique positive* ω*-periodic solution which is globally asymptotically stable.*

Proof. For any matrix norm $\|\cdot\|$ and any nonsingular matrix S, $\|\mathcal{K}\|_S=\|S^{-1}\mathcal{K}S\|$ also defines a matrix norm. Let $D=\operatorname{diag}(\xi_1,\xi_2,\ldots,\xi_n)$. Then the conditions (1) and (2) correspond to the column norms and Frobenius norm of matrix $D\mathcal{K}D^{-1}$, respectively. Condition (3) corresponds to the row norms of $D\mathcal{K}^TD^{-1}$, and note that $\rho(D\mathcal{K}^TD^{-1})=\rho(D\mathcal{K}D^{-1})$. Corollary 4.2 follows immediately. □

Remark 4.1. Taking $\xi_i=1, i=1,2,\ldots,n$, condition (1) reduces to the main results in Gopalsamy [9, 10]. Therefore, the previous results in [9, 10] are special cases of our results. In the next section, an example will be given to show that our results can be applied to that example while those of [9, 10] cannot be applied.

Now apply our results to the classical two-species Lotka–Volterra competition system which has been studied extensively in [3, 11, 20, 21, 22, 23]:

$$\begin{cases}\dot{y}_1(t) &= y_1(t)[b_1(t)-a_{11}(t)y_1(t)-a_{12}(t)y_2(t)],\\ \dot{y}_2(t) &= y_2(t)[b_2(t)-a_{21}(t)y_1(t)-a_{22}(t)y_2(t)].\end{cases}\tag{33}$$

Corollary 4.3. *Assume that the following conditions hold:*

(a) $m(a_{11})m(b_1)>m(a_{12})m(b_2)$, $m(a_{22})m(b_2)>m(a_{21})m(b_1)$*;*

(b) $\rho(\widetilde{K})<1$*, where*

$$\widetilde{K}=\begin{pmatrix}0 & \underline{a}_{22}^{-1}\overline{a}_{12}\\ \underline{a}_{11}^{-1}\overline{a}_{21} & 0\end{pmatrix}.$$

Then system (33) *has a unique positive* ω*-periodic solution which is globally asymptotically stable.*

Remark 4.2. By way of comparing our results, we first recall some previously known results in the literature [3, 11, 20, 21, 22, 23].

Theorem A. *Assume that the following condition holds:*

(I) $m(a_{11})m(b_1) > m(a_{12})m(b_2)\exp\{[m(b_2)+m(|b_2|)]\omega\}$ *and* $m(a_{22})m(b_2) > m(a_{21})m(b_1)\exp\{[m(b_1)+m(|b_1|)]\omega\}$.

Then system (33) *has at least one positive* ω*-periodic solution.*

THEOREM B. *In addition to* (I)*, further assume that*

(II) $\underline{a}_{11} > \overline{a}_{21}$ *and* $\underline{a}_{22} > \overline{a}_{12}$.

Then system (33) *has a unique positive* ω*-periodic solution which is globally asymptotically stable.*

We remark that conditions (a) and (b) in Corollary 4.3 are completely different from (I) and (II) in Theorems A and B. More specifically, it seems that we need condition (b) to guarantee the existence of a periodic solution, but actually, to guarantee the existence and global asymptotic stability of a unique periodic solution, conditions (a) and (b) in Corollary 4.3 are much weaker than (I) in Theorem A and (II) in Theorem B. It is in this sense that Corollary 4.3 generalizes and improves Theorems A and B. Therefore, our results are much different from the known results and thus essentially new.

Remark 4.3. The results in Gopalsamy [8, 20] require $\underline{a}_{11} > \overline{a}_{21}$ and $\underline{a}_{22} > \overline{a}_{12}$. That is, it is required that $\underline{a}_{11}^{-1}\overline{a}_{21} < 1$ and $\underline{a}_{22}^{-1}\overline{a}_{12} < 1$. We note that those conditions imply $\rho(\widetilde{K}) < 1$, where

$$\widetilde{K} = \begin{pmatrix} 0 & \underline{a}_{22}^{-1}\overline{a}_{12} \\ \underline{a}_{11}^{-1}\overline{a}_{21} & 0 \end{pmatrix}.$$

However, $\underline{a}_{11}^{-1}\overline{a}_{21} < 1$ and $\underline{a}_{22}^{-1}\overline{a}_{12} < 1$ cannot be inferred by $\rho(\widetilde{K}) < 1$. That is to say, there is a case: $\rho(\widetilde{K}) < 1$, but $\underline{a}_{11}^{-1}\overline{a}_{21}$ and $\underline{a}_{22}^{-1}\overline{a}_{12}$ may be bigger than 1. An example in the next section shows this fact. Therefore, our results also significantly improve the results in [8, 20].

Conclusions. (1) In this paper, we revisit the famous n-species Lotka–Volterra competitive system in a periodic environment. A set of new sufficient conditions are obtained to guarantee the existence and global asymptotic stability of a periodic solution in the multiple-species competition system. Our results are essentially new and much different from some previously known results. Moreover, applying our results to some special systems, we obtain some new criteria which generalize and improve the previously known results such as [3, 8, 9, 11, 20, 21, 22, 23].

(2) The main purpose of this paper is to propose a new methodology to study the Lotka–Volterra competitive system. The approaches used in this paper are based on Mawhin's coincidence degree, Lyapunov function, and matrix theory and its spectral theory. Mawhin's coincidence degree theory is extensively used to study the existence of periodic solutions. Note that different estimation techniques for the priori bounds of unknown solutions to the equation $Lx = \lambda Nx$ may lead to different results. There are many papers obtaining the priori bounds by employing the inequality $|x(t)| \leq |x(t_0)| + \int_0^\omega |\dot{x}(t)|dt$ (see, e.g., [11, 27, 28, 29, 30, 31, 32, 33, 34, 35]). To the best of our knowledge, no study has employed the matrix's spectral theory to obtain the priori bounds for biological systems so far. It is the first time that this new estimation technique for the priori bounds is employed to study global asymptotic stability and existence of periodic solution for the population dynamics. The method is much different from those in the previous references [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 27, 28, 29, 30, 31, 32, 33, 34, 35]. Due to this new method, our results are essentially new and very interesting.

5. Examples. In this section, some examples and their simulations are presented to illustrate the feasibility and effectiveness of our results.

Example 1. Consider the two-species competitive system

$$\left\{\begin{array}{lcl}\dot{x}_1(t) &=& x_1(t)[4-(3+\sin t)x_1(t)-\frac{1}{4}(1+\cos t)x_2(t)],\\ \dot{x}_2(t) &=& x_2(t)[8-(2+\sin t)x_1(t)-(3-\cos t)x_2(t)].\end{array}\right. \tag{34}$$

Corresponding to system (33), we have $\underline{a}_{11}=\underline{a}_{22}=2, \overline{b}_1=4, \overline{b}_2=8, \overline{a}_{12}=\frac{1}{2}, \overline{a}_{21}=3$. We see that $\underline{a}_{11}=2<\overline{a}_{21}=3$. This cannot meet the requirement $\underline{a}_{11}>\overline{a}_{21}$ and $\underline{a}_{22}>\overline{a}_{12}$ of the theorem in Gopalsamy [8]. Thus, the results obtained in [8] cannot be applied to this case.

However, it is not difficult to show that our results can be easily applied to system (34). In fact, on one hand, it is easy to check that system (34) satisfies condition (a) in Corollary 4.2. On the other hand, simple computation leads to

$$\widetilde{K}=\begin{pmatrix}0 & \frac{1}{2}\times\frac{1}{2}\\ \frac{1}{2}\times 3 & 0\end{pmatrix}=\begin{pmatrix}0 & \frac{1}{4}\\ \frac{3}{2} & 0\end{pmatrix}$$

and $\rho(\widetilde{K})=\frac{\sqrt{6}}{4}<1$. Thus, by Corollary 4.2, system (34) has a unique positive equilibrium $(\frac{8}{5},\frac{8}{5})$, which is globally asymptotically stable. Figure 1 shows the asymptotic behavior of system (34).

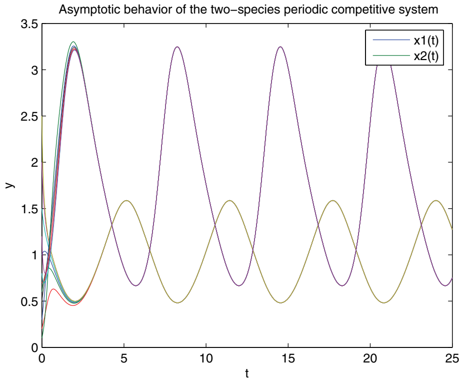


FIG. 1. *Asymptotic behavior of system* (34) *with initial values* $(x_1(0),x_2(0))=(0.2,0.1)$, $(0.6,0.3),(1,0.5),(1.5,0.7),(2,0.8),(2.5,1.2)$, *respectively,* $t\in[0,35]$.

Example 2. Consider the three-species competitive system

$$\left\{\begin{array}{lcl}\dot{x}_1(t) &=& x_1(t)[4-(3+\sin t)x_1(t)-\frac{1}{4}(1+\cos t)x_2(t)-\frac{1}{5}(1+\sin t)x_3(t)],\\ \dot{x}_2(t) &=& x_2(t)[4-\frac{1}{4}(1-\sin t)x_1(t)-(3+\sin t)x_2(t)],\\ \dot{x}_3(t) &=& x_3(t)[10-(2-\sin t)x_1(t)-\frac{1}{4}(1+\sin t)x_2(t)-(3+\cos t)x_3(t)].\end{array}\right. \tag{35}$$

Corresponding to system (3), we have $\underline{a}_{11}=\underline{a}_{22}=\underline{a}_{33}=2, \overline{b}_1=\overline{b}_2=4, \overline{b}_3=10, \overline{a}_{12}=\frac{1}{2}, \overline{a}_{13}=\frac{2}{5}, \overline{a}_{21}=\frac{1}{2}, \overline{a}_{23}=0, \overline{a}_{31}=3, \overline{a}_{32}=\frac{1}{2}$. We see that $\underline{a}_{11}=2<a_{21}+a_{31}=\frac{7}{2}$. This cannot meet the requirement $\underline{a}_{jj}>\sum_{i=1,i\neq j}^{3}\overline{a}_{ij}$ for $i=1,2,3$ of the theorem in [9, 10]. Thus, the results obtained in [4, 9, 10] cannot be applied to this case.

However, it is not difficult to show that our results can be easily applied to system

(35). In fact, simple computation leads to

$$\mathcal{K} = \begin{pmatrix} 0 & \underline{a}_{22}^{-1}\overline{a}_{12} & \underline{a}_{33}^{-1}\overline{a}_{13} \\ \underline{a}_{11}^{-1}\overline{a}_{21} & 0 & 0 \\ \underline{a}_{11}^{-1}\overline{a}_{31} & \underline{a}_{22}^{-1}\overline{a}_{32} & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2}\times\frac{1}{2} & \frac{1}{2}\times\frac{2}{5} \\ \frac{1}{2}\times\frac{1}{2} & 0 & 0 \\ \frac{1}{2}\times 3 & \frac{1}{2}\times\frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{4} & 0 & 0 \\ \frac{3}{2} & \frac{1}{4} & 0 \end{pmatrix}.$$

Hence, by using mathematica, we get

$$\rho(\mathcal{K}) = \max\ eigenvalues[\mathcal{K}] = 0.633982 < 1.$$

Thus, by Theorem 3.1, system (35) has a unique positive equilibrium which is globally asymptotically stable. Figure 2 shows the asymptotic behavior of system (35).

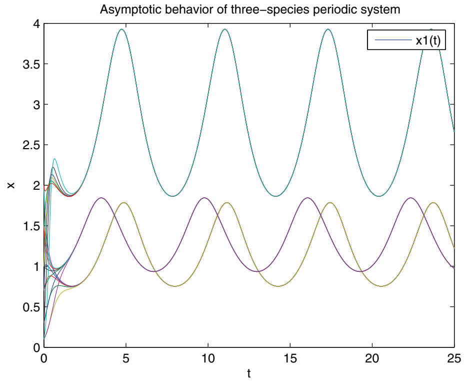


Fig. 2. *Asymptotic behavior of system* (35) *with initial values* $(x_1(0), x_2(0), y_1(0)) = (0.1, 0.1, 0.1), (0.3, 0.3, 0.3), (0.7, 0.7, 0.7), (1, 1, 1), (1.5, 1.5, 1.5), (2, 2, 2)$*, respectively,* $t \in [0, 35]$.

Remark 5.1. In this example, one can observe that though the spectral $\rho(\mathcal{K}) < 1$, the matrix norms (including the row norm, the column norm, and the Frobenius norm) of matrix $\mathcal{K}$ are all bigger than 1. For instance, the column norm

$$\|\mathcal{K}\|_1 = \max_{1\le j\le 3}\left\{\underline{a}_{jj}^{-1}\sum_{i=1, i\ne j}^{3}\overline{a}_{ij}\right\} = 0 + \frac{3}{2} + \frac{1}{4} > 1.$$

This fact implies that our results are more general than those in [8, 9, 10].

Acknowledgments. The authors are very grateful to the editor and referees for their careful reading of the manuscript and a number of excellent suggestions which improved the presentation of this paper.

REFERENCES

[1] L. Chen, *Mathematical Models and Methods in Ecology,* Science Press, Beijing, 1988 (in Chinese).

[2] P. de Mottoni and A. Schiaffino, *Competition system with periodic coefficients: A geometric approach*, J. Math. Biol., 11 (1981), pp. 319–335.

[3] J. M. Cushing, *Two species competition in a periodic environment*, J. Math. Biol., 10 (1980), pp. 385–400.

[4] J. M. Cushing, *Periodic Lotka–Volterra competition equations*, J. Math. Biol., 24 (1986), pp. 381–403.

[5] S. Ahmad, *Convergence and ultimate bounds of solutions of nonautonomous Volterra-Lotka competition equations*, J. Math. Anal. Appl., 127 (1987), pp. 377–387.

[6] S. Ahmad, *On almost periodic solutions of the competiting species problems*, Proc. Amer. Math. Soc., 102 (1988), pp. 855–861.

[7] S. Ahmad, *On nonautonomous Volterra-Lotka competition equations*, Proc. Amer. Math. Soc., 177 (1993), pp. 199–204.

[8] K. Gopalsamy, *Exchange of equilibria in two species Lotka–Volterra competition models*, J. Austral. Math. Soc. Ser. B, 24 (1982), pp. 160–170.

[9] K. Gopalsamy, *Global asymptotic stability in a periodic Lotka–Volterra system*, J. Austral. Math. Soc. Ser. B, 27 (1985), pp. 66–72.

[10] K. Gopalsamy, *Global asymptotic stability in Volterra's population systems*, J. Math. Biol., 19 (1984), pp. 157–168.

[11] C. Alvarez and A. C. Lazer, *An application of topological degree to the periodic competing species problem*, J. Austral. Math. Soc. Ser. B, 28 (1986), pp. 202–219.

[12] A. Tineo and C. Alvarez, *A different consideration about the globally asymptotically stable solution of the periodic n-competing species problem*, J. Math. Anal. Appl., 159 (1991), pp. 44–45.

[13] H. P. Zhu, S. A. Campbell, and G. S. K. Wolkowicz, *Bifurcation analysis of a predator-prey system with nonmonotonic function response*, SIAM J. Appl. Math., 63 (2002), pp. 636–682.

[14] D. Xiao and H. P. Zhu, *Multiple focus and Hopf bifurcations in a predator-prey system with nonmonotonic functional response*, SIAM J. Appl. Math., 66 (2006), pp. 802–819.

[15] Y. H. Xia, F. D. Chen, A. Chen, and J. Cao, *Existence and global attractivity of an almost periodic ecological model*, Appl. Math. Comput., 157 (2004), pp. 449–475.

[16] Y. H. Xia and J. Cao, *Almost periodic solutions for an ecological model with infinite delays*, Proc. Edinb. Math. Soc., 50 (2007), pp. 229–249.

[17] Y. H. Xia and J. Cao, *Global attractivity of a periodic ecological model with m-predators and n-preys by "pure-delay type" system*, Comput. Math. Appl., 52 (2006), pp. 829–852.

[18] Y. H. Xia, J. Cao, and M. Han, *A new analytical method for the linearization of dynamic equation on measure chains*, J. Differential Equations, 235 (2007), pp. 527–543.

[19] Y. H. Xia, M. Han, and W. Ding, *New conditions on the global asymptotic stability of equilibrium in Lotka–Volterra's population systems*, Comm. Pure Appl. Anal., to appear.

[20] K. Gopalsamy, *Stability and Oscillation in Delay Differential Equations of Population Dynamics*, Mathematics and its Applications 74, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[21] A. Shibata and N. Satiô, *Time delays and chaos in two competition systems*, Math. Biosci., 51 (1980), pp. 199–211.

[22] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Mathematics in Science and Engineering 191, Academic Press, Boston, 1993.

[23] B. S. Goh, *Management and Analysis of Biological Populations*, Elsevier Scientific, Dordrecht, The Netherlands, 1980.

[24] I. Barbălat, *Systems d'equations differentielle d'oscillations nonlineaires*, Rev. Roumaine Math. Pures. Appl., 4 (1959), pp. 267–270.

[25] J. P. Lasalle, *The Stability of Dynamical Systems*, SIAM, Philadelphia, 1976.

[26] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[27] M. Fan and Q. Wang, *Periodic solutions of a class of nonautonomous discrete time semi-ratio-dependent predator-prey systems*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 563–574.

[28] M. Fan and K. Wang, *Global periodic solutions of a generalized n-species Gilpin-Ayala competition model*, Comput. Math. Appl., 40 (2000), pp. 1141–1151.

[29] M. Fan and K. Wang, *Existence and global attractivity of positive periodic solutions of periodic n-species Lotka–Volterra competition systems with several deviating arguments*, Math. Biosci., 160 (1999), pp. 47–61.

[30] M. Fan, Q. Wang, and X. F. Zou, *Dynamics of a nonautonomous ratio-dependent predator-prey system*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 97–118.

[31] X. F. Zou and X.-H. Tang, 3/2*-type criteria for global attractivity of Lotka–Volterra competition system without instantaneous negative feedbacks*, J. Differential Equations, 186 (2002), pp. 420–439.

[32] Y. H. Xia, J. Cao, and S. S. Cheng, *Periodicity in a Lotka–Volterra mutualism system with several delays*, Appl. Math. Modeling, 31 (2007), pp. 1960–1969.

[33] Y. H. Xia and M. Han, *Multiple Periodic Solutions of a Ratio-Dependent Predator-Prey*

Model, Chaos, Solitons Fractals, to appear.

[34] R. E. Gaines and J. L. Mawhin, *Coincidence Degree and Nonlinear Differential Equations*, Springer-Verlag, Berlin-New York, 1977.

[35] D. Guo, J. Sun, and Z. Liu, *Functional Method in Nonlinear Ordinary Differential Equations*, Shangdong Scientific Press, Shandong, 2005.

[36] J. Li, Y. C. Zhou, Z. Ma, and J. M. Hyman, *Epidemiological models for mutating pathogens*, SIAM J. Appl. Math., 65 (2004), pp. 1–23.

[37] J. M. Hyman and J. Li, *Behavior changes in SIS STD models with selective mixing*, SIAM J. Appl. Math., 57 (1997), pp. 1082–1094.

[38] C. Castillo-Chavez, W. Z. Huang, and J. Li, *Competitive exclusion and coexistence of multiple strains in an SIS STD model*, SIAM J. Appl. Math., 59 (1999), pp. 1790–1811.

[39] C. Castillo-Chavez, W. Z. Huang, and J. Li, *Competitive exclusion in gonorrhea models and other sexually transmitted diseases*, SIAM J. Appl. Math., 56 (1996), pp. 494–508.

[40] J. K. Hale, *Theory of Functional Differential Equations*, 2nd ed., Springer-Verlag, New York-Heidelberg, 1977.

[41] T. A. Burton, *Stability and Periodic Solution of Ordinary and Functional-Differential Equations*, Academic Press, Orlando, FL, 1985.

[42] H. Minc, *Nonnegative Matrices*, John Wiley and Sons, New York, 1988

[43] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1598–1617

BOUNDS FOR THE EFFECTIVE STRESS OF CLASSICAL AND STRAIN GRADIENT PLASTIC COMPOSITES*

VIET HA HOANG†

Abstract. Given an average strain, rigorous bounds are established for the stress in a deformation of a plastic composite material, which follows a power law. The deformation theory of strain gradient plasticity, which introduces an internal material length scale, is used. It falls into the classical deformation theory of elasto-plasticity when this length scale equals zero. The method employs the idea by Milton and Serkov [*J. Mech. Phys. Solids*, 48 (2000), pp. 1259–1324] and other techniques for bounding effective energy. We derive two stress bounds which closely relate to the Reuss lower bound and the Hashin–Shtrikman upper bound for the energy. We then study numerically the dependence on the internal length scale of the magnitude of the stress and the region in the stress space determined by these two bounds in which the macro stress must lie. The results confirm the prediction made by Fleck and Willis [*J. Mech. Phys. Solids*, 52 (2004), pp. 1855–1888] for the macroscopic uniaxial response by differentiating their energy bounds.

Key words. stress bounds, plastic composites, strain gradient, compensated compactness, null-Lagrangian

AMS subject classifications. 74Q20, 74C05, 74E30

DOI. 10.1137/070700206

1. Introduction.

For nonlinear composite materials, a correct bound for the constitutive relation (e.g., a bound for the average stress in terms of the average strain) may not be induced from an energy bound: Differentiating an energy bound with respect to the average strain may not give a bound for the average stress. A separate technique for bounding the constitutive relations needs to be developed. Milton and Serkov [8] propose an approach for bounding the current in nonlinear conducting composites. They present it as an extension of earlier works on bounding the yield surface of plastic composites (Kohn and Little [5] and Nesi et al. [10]) which are applications of the compensated compactness due to Tartar [20, 21] and Murat and Tartar [9] and also the translation method of Lurie and Cherkaev [6, 7] and Murat and Tartar [9]. The key idea is to use a functional of the current $j(x)$ and the electric field $e(x)$ which equals 0 in the admissible range of (j, e) and equals infinity elsewhere, and to employ a translated function $Q(j, e)$ such that $Q(j(x), e(x))$ is quasiconvex. For simplicity, they use $Q = j.e$ which is a null Lagrangian, but predict that better bounds may be obtained by using other functions. The method is further developed for nonlinear elastic composites by Talbot and Willis [19], using a homogeneous comparison linear material, and by Peigney [11].

In this paper, we study bounds for the effective stress of a power law plastic composite under a uniform boundary strain. We use the strain gradient plasticity theory of Fleck and Willis [3] which introduces an internal material length scale. This is a modification of the phenomenological constitutive law by Fleck and Hutchinson [2]. It reduces to classical deformation theory of elasto-plasticity when the internal

*Received by the editors August 15, 2007; accepted for publication (in revised form) December 12, 2008; published electronically March 18, 2009. This work was supported by the Meggitt Fellowship awarded by Emmanuel College, Cambridge.
http://www.siam.org/journals/siap/69-6/70020.html

†Emmanuel College, Cambridge, CB2 3AP, England, and Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WA, England (V.H.Hoang@damtp.cam.ac.uk).

length scale is 0; thus this work holds for classical plasticity. We cannot apply directly the techniques of [8] and [19] as the stress is not a function of the strain. It depends nonlinearly on the plastic strain and its derivatives. The approach we develop is an adaptation of that of Milton and Serkov [8], and the ideas for bounding the effective energy of nonlinear composites.

In the next section, we introduce the strain gradient plasticity theory. A bound that resembles the Reuss bound for the energy is deduced in section 3, which gives a lower bound for the stress magnitude. This lower bound, indeed, can be deduced by differentiating the Reuss lower bound for the effective energy in Fleck and Willis [3]. A stress bound which closely relates to the Hashin–Shtrikman upper bound is found in section 4. For the stress magnitude, deep in the plastic range, it gives a bound which can be obtained by differentiating the Hashin–Shtrikman upper energy bound in [3]. This is expected as deep in the plastic range, the upper energy bound in [3] follows a power law, and given that the exact effective energy also follows a power law, the energy bound in effect gives a bound for the reference strength. However, for smaller total strain, the energy upper bound does not follow a power law; differentiating the energy bound does not give a rigorous bound for the magnitude of the stress. Our results provide rigorous stress bounds for the whole range of the total strain. We present some numerical results in section 5. They show that the Hashin–Shtrikman type bound for the magnitude of the stress increases with the material length scale. This is in agreement with the prediction Fleck and Willis [3] made by differentiating the upper energy bound. Deep in the plastic regime, the upper bound for the magnitude of the stress follows a power law, which is also in agreement with [3]. For a specific boundary strain, our bounds give a particular region in the stress space in which the stress tensor lies. We give and analyze some examples for a couple of particular boundary strains, which show that the region for a larger material length scale contains the region for smaller ones. This agrees with the previously mentioned fact that bounds for the stress magnitude increase with the material length scale. The paper ends with a short concluding section 6 in which the main results are summarized.

Before Fleck and Willis [3], bounds and estimates for the overall properties of plastic composites had been studied in several works. A number of them, e.g., Ponte Castañeda [12], Ponte Castañeda and De Botton [13], and Suquet [17], study conventional plastic materials where bounds and estimates for the effective energy density, and thus for the yield strength, are deduced. Extending these works, composites, whose components are under the effect of strain gradient which introduces internal length scales, are considered in Smyshlyaev and Fleck [14, 15, 16]. The present paper complements these works by contributing rigorous bounds for the effective stress.

2. Strain gradient energy.

We consider a plastic composite material whose energy functional follows the strain gradient deformation theory of Fleck and Willis [3], which modifies the original theory of Fleck and Hutchinson [2]. We assume that the material is incompressible and is elastically homogeneous. The energy density function is defined as

$$U(\varepsilon_{ij}, \varepsilon^P_{ij}, \varepsilon^P_{ij,k}) = \mu(\varepsilon_{ij} - \varepsilon^P_{ij})(\varepsilon_{ij} - \varepsilon^P_{ij}) + V(\varepsilon^P_{ij}, \varepsilon^P_{ij,k}),$$

where the potential $V = V(E_P)$; E_P is the effective plastic strain measure defined by

$$E_P = \sqrt{\frac{2}{3}}(\varepsilon^P_{ij}\varepsilon^P_{ij} + l^2\varepsilon^P_{ij,k}\varepsilon^P_{ij,k})^{1/2}.$$

The underlying idea is that $\varepsilon^P_{ij}\varepsilon^P_{ij}$ provides a measure of the density of the statistically stored dislocations while $\varepsilon^P_{ij,k}\varepsilon^P_{ij,k}$ provides a measure of the geometrically necessary dislocations. The original theory of Fleck and Hutchinson [2] contains three material scales l_1, l_2, and l_3. Here following [3], we consider a simplified version which contains a single scale l. The material is a composite so that the potential V depends on the position x and is rapidly oscillating. We restrict our attention to the case of power law components; i.e., the potential V has the form

$$V(x,\varepsilon^P_{ij},\varepsilon^P_{ij,k})=\frac{\Sigma(x)e_0}{N+1}\Big(\frac{E_P}{e_0}\Big)^{N+1}, \tag{2.1}$$

where $\Sigma(x)$ is the reference strength, and $0<N<1$. The strain ε_{ij} is defined as usual as

$$\varepsilon_{ij}=\frac{1}{2}\Big(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\Big).$$

In this setting, the displacement u_i and the plastic strain ε^P_{ij} are treated as dependent variables on equal footing. Assuming that an affine displacement $u_i=\bar\varepsilon_{ij}x_j$ ($\bar\varepsilon$ is a constant tensor) is prescribed on the boundary of a domain $\Omega\in\mathbb{R}^3$ which has unit volume, the deformation minimizes the energy functional

$$\Psi(u_i,\varepsilon^P_{ij})=\int_\Omega U(\varepsilon_{ij},\varepsilon^P_{ij},\varepsilon^P_{ij,k})dx. \tag{2.2}$$

We introduce the conjugate variables

$$\begin{aligned}\sigma_{ij}&=\frac{\partial U}{\partial\varepsilon_{ij}}=2\mu(\varepsilon_{ij}-\varepsilon^P_{ij}),\\ s_{ij}&=\frac{\partial U}{\partial\varepsilon^P_{ij}}=-2\mu(\varepsilon_{ij}-\varepsilon^P_{ij})+\frac{\partial V}{\partial\varepsilon^P_{ij}},\\ \tau_{ijk}&=\frac{\partial U}{\partial\varepsilon^P_{ij,k}}=\frac{\partial V}{\partial\varepsilon^P_{ij,k}}.\end{aligned} \tag{2.3}$$

Setting to zero the first variation of (2.2), we get the equilibrium equations

$$\begin{aligned}\sigma_{ij,j}&=p_{,i},\\ \sigma_{ij}+\tau_{ijk,k}&=\frac{\partial V}{\partial\varepsilon^P_{ij}},\\ \tau_{ijk}n_k&=0,\end{aligned} \tag{2.4}$$

where p is a pressure field whose presence is due to the incompressibility. From this we have

$$\tau_{ijk,k}=s_{ij}, \tag{2.5}$$

and

$$\sigma_{ij}=\frac{\partial V}{\partial\varepsilon^P_{ij}}-\Big(\frac{\partial V}{\partial\varepsilon^P_{ij,k}}\Big)_{,k}. \tag{2.6}$$

As the constitutive materials are incompressible, $\varepsilon_{ii}=\varepsilon^P_{ii}=0$. The phases are assumed to be perfectly bonded together so that there is no energy built up on the surface between different components as considered in, e.g., Gudmundson [4] and Aifantis and Willis [1]. Throughout, $\langle.\rangle$ denotes the mean value of a function in Ω; we denote $\bar\sigma=\langle\sigma\rangle$, $\bar\varepsilon=\langle\varepsilon\rangle$, and $\bar\varepsilon^P=\langle\varepsilon^P\rangle$. As usual, repeated indices indicate summation.

3. Reuss-type bound for the stress. In this section, we adopt the approach by Milton and Serkov [8] to find a Reuss-type bound for the stress, using the null Lagrangian $\sigma : \varepsilon$. However, unlike the situations in [8], [19], and [11], here we need to take into account both the strain fields ε and ε^P, and the stress σ is a complicated nonlinear function of ε^P and its derivatives. For a constant tensor λ, it is not possible to deal with $\sup_\varepsilon\{\lambda : \varepsilon - \sigma : \varepsilon/2\}$ as in [8]. We, instead, work with

$$F(\lambda) = \sup\left\{\int_\Omega \left(\lambda : \varepsilon^P - \frac{1}{2}\sigma : \varepsilon^P\right) dx\right\},$$

where the supremum is taken with respect to all the tensor fields $\varepsilon^P(x)$ and $\sigma(x)$ that satisfy conditions (2.6), (2.3c), and (2.4c). From these conditions

$$\begin{aligned}\int_\Omega \sigma_{ij}\varepsilon^P_{ij} dx &= \int_\Omega \frac{\partial V}{\partial \varepsilon^P_{ij}}\varepsilon^P_{ij} dx + \int_\Omega \frac{\partial V}{\partial \varepsilon^P_{ij,k}}\varepsilon^P_{ij,k} dx\\ &= \int_\Omega \Sigma(x) e_0^{-N}\left(\frac{2}{3}\right)^{\frac{N+1}{2}} (\varepsilon^P : \varepsilon^P + l^2\nabla\varepsilon^P : \nabla\varepsilon^P)^{\frac{N+1}{2}} dx.\end{aligned} \tag{3.1}$$

Therefore

$$F(\lambda) \le \sup_{\varepsilon^P}\left\{\int_\Omega \left(\lambda : \varepsilon^P - \frac{\Sigma(x)e_0^{-N}}{2}\left(\frac{2}{3}\right)^{\frac{N+1}{2}} (\varepsilon^P : \varepsilon^P + l^2\nabla\varepsilon^P : \nabla\varepsilon^P)^{\frac{N+1}{2}}\right) dx\right\}, \tag{3.2}$$

where the supremum is taken over all the tensor fields $\varepsilon^P_{ij}(x)$. Since

$$\begin{aligned}\lambda : \varepsilon^P - \frac{\Sigma(x)e_0^{-N}}{2}\left(\frac{2}{3}\right)^{\frac{N+1}{2}} &(\varepsilon^P : \varepsilon^P + l^2\nabla\varepsilon^P : \nabla\varepsilon^P)^{\frac{N+1}{2}}\\ &\le \sup_{\varepsilon^P}\left\{\lambda : \varepsilon^P - \frac{\Sigma(x)e_0^{-N}}{2}\left(\frac{2}{3}\varepsilon^P : \varepsilon^P\right)^{\frac{N+1}{2}}\right\}\\ = Ne_0 2^{1/N}(\Sigma(x))^{-1/N}&\left(\frac{2}{3}\right)^{-\frac{N+1}{2N}}(N+1)^{-\frac{N+1}{N}}(\lambda : \lambda)^{\frac{N+1}{2N}},\end{aligned}$$

we have

$$F(\lambda) \le Ne_0 2^{1/N}\langle\Sigma^{-1/N}\rangle\left(\frac{2}{3}\right)^{-\frac{N+1}{2N}}(N+1)^{-\frac{N+1}{N}}(\lambda : \lambda)^{\frac{N+1}{2N}}.$$

From

$$\int_\Omega (\lambda : \varepsilon^P - \frac{1}{2}\sigma : \varepsilon^P) dx - F(\lambda) \le 0,$$

we deduce

$$\begin{aligned}\int_\Omega (\lambda : \varepsilon - \frac{1}{2}\sigma : \varepsilon)dx - F(\lambda) &\le \int_\Omega (\lambda(\varepsilon - \varepsilon^P) - \frac{1}{2}\sigma : (\varepsilon - \varepsilon^P))dx\\ &= \int_\Omega \Big(\frac{1}{2\mu}\lambda : \sigma - \frac{1}{4\mu}\sigma : \sigma\Big) dx.\end{aligned}$$

Hence

$$\lambda : \bar\varepsilon - \frac{1}{2}\bar\sigma : \bar\varepsilon - F(\lambda) \le \frac{1}{2\mu}\lambda : \bar\sigma - \frac{1}{4\mu}\bar\sigma : \bar\sigma$$

(from the Cauchy–Schwarz inequality, $\int_\Omega \sigma : \sigma dx \geq \bar{\sigma} : \bar{\sigma}$). Thus

$$\lambda : \Big(\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}\Big) - Ne_0 2^{1/N} \langle \Sigma^{-1/N} \rangle \Big(\frac{2}{3}\Big)^{-\frac{N+1}{2N}} (N+1)^{-\frac{N+1}{N}} (\lambda : \lambda)^{\frac{N+1}{2N}} \leq \frac{1}{2}\bar{\sigma} : \Big(\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}\Big).$$

Optimizing over the field λ, the optimal value of the left-hand side is obtained when

$$\lambda : (\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}) = e_0 2^{1/N} \langle \Sigma^{-1/N} \rangle \Big(\frac{2}{3}\Big)^{-\frac{N+1}{2N}} (N+1)^{\frac{-1}{N}} (\lambda : \lambda)^{\frac{N+1}{2N}}.$$

We then obtain the bound

$$\Big(\Big(\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}\Big) : \Big(\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}\Big)\Big)^{\frac{1+N}{2}} \Big(\frac{2}{3}\Big)^{\frac{N+1}{2}} e_0^{-N} \langle \Sigma^{-1/N} \rangle^{-N} \leq \bar{\sigma} : \Big(\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}\Big). \tag{3.3}$$

From

$$\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu} = \bar{\varepsilon}^P,$$

the bound can be written as

$$(\bar{\varepsilon}^P : \bar{\varepsilon}^P)^{\frac{1+N}{2}} \Big(\frac{2}{3}\Big)^{\frac{N+1}{2}} e_0^{-N} \langle \Sigma^{-1/N} \rangle^{-N} \leq \bar{\sigma} : \bar{\varepsilon}^P.$$

Using the Cauchy–Schwarz inequality for the right-hand side

$$\bar{\sigma} : \bar{\varepsilon}^P \leq (\bar{\sigma} : \bar{\sigma})^{1/2} (\bar{\varepsilon}^P : \bar{\varepsilon}^P)^{1/2}, \tag{3.4}$$

we deduce the following bound for $\|\bar{\sigma}\|$

$$\|\bar{\varepsilon}^P\|^N \Big(\frac{2}{3}\Big)^{\frac{N+1}{2}} e_0^{-N} \langle \Sigma^{-1/N} \rangle^{-N} \leq \|\bar{\sigma}\|.$$

This can be obtained by differentiating the Reuss bound for the effective energy (3.25) of Fleck and Willis [3] with respect to $\bar{\varepsilon}^P_{ij}$; i.e., the left-hand side of the above equals

$$\Big(\frac{\partial V_R}{\partial \bar{\varepsilon}^P_{ij}} . \frac{\partial V_R}{\partial \bar{\varepsilon}^P_{ij}}\Big)^{1/2},$$

where $V_R(\bar{\varepsilon}^P_{ij})$ is the Fleck and Willis's Reuss bound and $\bar{\varepsilon}^P$ stands for ε^{P0} in Fleck and Willis's notations.

4. Hashin–Shtrikman-type bound for the stress. We now deduce a Hashin–Shtrikman-type bound for the constitutive relation. The idea follows closely those for bounding the effective energy of a nonlinear composite and those of Milton–Serkov [8] and Talbot and Willis [19]. We use the null Lagrangian $\sigma : \varepsilon$. As in the previous section, it is necessary to manipulate the quantity $\int_\Omega \sigma : \varepsilon^P dx$. We will write this in terms of σ, s, and τ rather than ε^P and $\nabla\varepsilon^P$; the reason will be given later. From (2.3a,b) and (3.1),

$$\int_\Omega \sigma : \varepsilon^P dx = e_0 \Big(\frac{2}{3}\Big)^{-\frac{N+1}{2N}} \int_\Omega \Sigma(x)^{-1/N} \Big((\sigma + s) : (\sigma + s) + \frac{1}{l^2}\tau : \tau\Big)^{\frac{N+1}{2N}} dx. \tag{4.1}$$

The expression under the integral on the right-hand side is nonlinear; thus following the well-known procedure for bounding nonlinear composites (see, for example, [3] among many other references), we choose a "linear comparison" quantity

$$b(x)\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big).$$

Indeed, in the unrealistic case of linear components, i.e., $N=1$,

$$\int_\Omega \sigma:\varepsilon^P dx=\int_\Omega b(x)\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)dx,$$

where $b(x)=(3/2)e_0\Sigma(x)^{-1}$. We then choose a homogeneous comparison material; i.e., we compare this quantity to $b_0((\sigma+s):(\sigma+s)+\tau:\tau/l^2)$ where $b_0\le b(x)$. This again follows the usual procedure for bounding the effective energy, and for bounding the constitutive relation of Talbot and Willis [19].

For each tensor β, we define

$$\begin{aligned}F(\beta)&=\sup_\rho\{\beta:\rho+(b_0-b(x))(\rho:\rho)\}\\&=\frac{\beta:\beta}{4(b(x)-b_0)}.\end{aligned}$$

For $\rho=\sigma+s$,

$$\beta:(\sigma+s)+(b_0-b(x))((\sigma+s):(\sigma+s))-\frac{\beta:\beta}{4(b(x)-b_0)}\le 0,$$

so

$$\beta:(\sigma+s)+(b_0-b(x))\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)-\frac{\beta:\beta}{4(b(x)-b_0)}\le 0.$$

From this

$$\begin{gathered}\beta:(\sigma+s)+(b_0-b(x))\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)\\ \frac{-\beta:\beta}{4(b(x)-b_0)}+\frac{1}{4\mu}\sigma:\sigma-\frac{1}{2}\sigma:(\varepsilon-\varepsilon^P)\le 0.\end{gathered}$$

Taking the integral over Ω,

$$\begin{aligned}&-\frac{1}{2}\bar\sigma:\bar\varepsilon+\int_\Omega\Bigg(\frac{1}{4\mu}\sigma:\sigma+\beta:(\sigma+s)+b_0\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)\\&\qquad-\frac{\beta:\beta}{4(b(x)-b_0)}\Bigg)dx+\int_\Omega\Bigg(\frac{1}{2}\sigma:\varepsilon^P-b(x)\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)\Bigg)dx\le 0.\end{aligned}\tag{4.2}$$

Let K be the set of triplets of tensors (σ,s,τ) that satisfy $\sigma_{ij,j}=p_{,i}$ for a scalar function $p(x)$, $\tau_{ijk}n_k=0$ on $\partial\Omega$ and $\tau_{ijk,k}=s_{ij}$. We proceed by taking the infimum of the left-hand side of (4.2) with respect to $(\sigma,s,\tau)\in K$. From (4.1),

$$\begin{aligned}&\int_\Omega\frac{1}{2}\sigma:\varepsilon^P-b(x)\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)dx\\&\ge\int_\Omega\inf_Z\left\{\left(\frac{1}{2}e_0\left(\frac{2}{3}\right)^{-\frac{N+1}{2N}}(\Sigma(x))^{-1/N}Z^{(N+1)/(2N)}-b(x)Z\right)\right\}dx,\end{aligned}\tag{4.3}$$

where Z is a scalar quantity. Computing the infimum, we get

$$\begin{aligned}&\inf_{\{\sigma,s,\tau\}\in K}\int_\Omega \frac{1}{2}\sigma : \varepsilon^P - b(x)\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)dx\\ &\geq -\frac{1-N}{2N}\Big(\frac{1+N}{2N}\Big)^{\frac{N+1}{N-1}} 2^{\frac{2N}{1-N}} e_0^{\frac{2N}{N-1}} \Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}} \int_\Omega \Sigma(x)^{\frac{2}{1-N}} b(x)^{\frac{1+N}{1-N}} dx.\end{aligned}$$

Inequality (4.3) bears some resemblance to the procedure for bounding the effective energy of a nonlinear composite initiated by Ponte-Castañeda [12]. Note that it is not possible to obtain a finite infimum if we use (3.1) instead of (4.1) and a linear comparison material.

Now we find

$$\inf_{\{\sigma,s,\tau\}\in K}\int_\Omega \Big(\frac{1}{4\mu}\sigma:\sigma+\beta:(\sigma+s)+b_0\Big((\sigma+s):(\sigma+s)+\frac{1}{l^2}\tau:\tau\Big)\Big)dx.$$

Taking a variation with respect to σ, we find that for all stress tensors σ' with zero mean that satisfy (2.4a) and $\sigma'_{ii}=0$, we have

$$\int\Big(\beta+\frac{1}{2\mu}\sigma+2b_0(\sigma+s)\Big):\sigma' dx = 0. \tag{4.4}$$

Substituting $s_{ij}=\tau_{ijk,k}$ and taking a variation of τ, for all fields τ' such that $\tau'_{ijk}n_k = 0$ on $\partial\Omega$, we have

$$\int_\Omega\Big(-\beta_{ij,k}-2b_0(\tau_{ijl,lk}+\sigma_{ij,k})+\frac{2b_0}{l^2}\tau_{ijk}\Big)\tau'_{ijk}dx = 0. \tag{4.5}$$

Equation (4.4) implies that

$$\begin{aligned}\beta_{ij}+\frac{1}{2\mu}\sigma_{ij}+2b_0(\sigma_{ij}+s_{ij}) &= \frac{1}{2}(v_{i,j}+v_{j,i}),\\ v_{i,i} &= 0\end{aligned} \tag{4.6}$$

for a vector field $v(x)\in\mathbb{R}^3$ with an affine boundary displacement, i.e., $v_i=\bar{e}_{ij}x_j$ on the boundary $\partial\Omega$; so from (2.4a)

$$v_{i,jj}-(4b_0+1/\mu)p_{,i} = 2\beta_{ij,j}+4b_0 s_{ij,j}, \tag{4.7}$$

the constant tensor $\bar{e}$ satisfies

$$\bar{e}_{ij} = \langle\beta_{ij}\rangle+\frac{\bar{\sigma}_{ij}}{2\mu}+2b_0\bar{\sigma}_{ij}. \tag{4.8}$$

Choosing $\tau'_{ijk}(x)=\theta_{ij,k}(x)$ where θ satisfies the condition $\theta_{ij,k}n_k=0$ on $\partial\Omega$, we get

$$\begin{aligned}&\int_\Omega\Big(\beta_{ij,k}+2b_0(\tau_{ijl,lk}+\sigma_{ij,k})-\frac{2b_0}{l^2}\tau_{ijk}\Big)_{,k}\theta_{ij}\\ &+\int_{\partial\Omega}(\beta_{ij,k}+2b_0(\tau_{ijl,lk}+\sigma_{ij,k})-\frac{2b_0}{l^2}\tau_{ijk}\Big)n_k\theta_{ij}dS = 0.\end{aligned}$$

As $\theta(x)$ can be chosen arbitrarily, we deduce that

$$\Big(\beta_{ij,k}+2b_0(\tau_{ijl,lk}+\sigma_{ij,k})-\frac{2b_0}{l^2}\tau_{ijk}\Big)_{,k} = 0,$$

and

$$\Big(\beta_{ij,k} + 2b_0(\tau_{ijl,lk} + \sigma_{ij,k}) - \frac{2b_0}{l^2}\tau_{ijk}\Big)n_k = 0$$

on $\partial\Omega$. From these

$$\beta_{ij,kk} + 2b_0(s_{ij,kk} + \sigma_{ij,kk}) - \frac{2b_0}{l^2}s_{ij} = 0, \tag{4.9}$$

with the boundary condition

$$\Big(\beta_{ij,k} + 2b_0(\sigma_{ij,k} + s_{ij,k})\Big)n_k = 0.$$

Using $\sigma' = \sigma - \bar{\sigma}$ and $\tau' = \tau$ in (4.4) and (4.5), we deduce that

$$\begin{aligned}\inf_{\{\sigma,s,\tau\}\in K}\int_\Omega\bigg(\frac{1}{4\mu}\sigma : \sigma + \beta : (\sigma + s) + b_0\Big((\sigma + s) : (\sigma + s) + \frac{1}{l^2}\tau : \tau\Big)\bigg)dx \\ = \frac{1}{2}\int \beta : (\sigma + s)dx + \frac{1}{2}\langle\beta\rangle : \bar{\sigma} + \frac{1}{4\mu}\bar{\sigma} : \bar{\sigma} + b_0\bar{\sigma} : \bar{\sigma}.\end{aligned}$$

We then deduce the following bound for the stress:

$$\begin{aligned}\int_\Omega\bigg(\frac{1}{2}\beta : (\sigma + s) - \frac{\beta : \beta}{4(b(x) - b_0)}\bigg)dx + \frac{1}{2}\langle\beta\rangle : \bar{\sigma} + \frac{1}{4\mu}\bar{\sigma} : \bar{\sigma} + b_0\bar{\sigma} : \bar{\sigma} - \frac{1}{2}\bar{\sigma} : \bar{\varepsilon} \\ - \frac{1 - N}{2N}\Big(\frac{1 + N}{2N}\Big)^{\frac{N+1}{N-1}} 2^{\frac{2N}{1-N}} e_0^{\frac{2N}{N-1}}\Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}}\int_\Omega \Sigma(x)^{\frac{2}{1-N}} b(x)^{\frac{1+N}{1-N}}dx \le 0.\end{aligned} \tag{4.10}$$

Solving (4.7) and (4.9), we find that σ and s depend on β linearly. In particular, there are fourth-order tensors $M_{ijkl}(x)$ and $N_{ijkl}(x)$ such that

$$\sigma(x) - \bar{\sigma} = \int M_{ijkl}(x - x')(\beta(x') - \langle\beta\rangle)dx', \quad s(x) = \int N_{ijkl}(x - x')(\beta(x') - \langle\beta\rangle)dx',$$

the integral being taken over the infinite domain due to the small correlation of the phases. We deduce this in Appendix A. Optimizing (4.10) by taking the supremum of

$$\int\bigg(\frac{1}{2}\beta : (\sigma + s) - \frac{\beta : \beta}{4(b(x) - b_0)}\bigg)dx + \frac{1}{2}\langle\beta\rangle : \bar{\sigma}, \tag{4.11}$$

with respect to β, at the value of β that the supremum is attained, the bound becomes

$$\begin{aligned}&\frac{1}{2}\langle\beta\rangle : \bar{\sigma} + \frac{1}{4\mu}\bar{\sigma} : \bar{\sigma} + b_0\bar{\sigma} : \bar{\sigma} - \frac{1}{2}\bar{\sigma} : \bar{\varepsilon} \\ &-\frac{1 - N}{2N}\Big(\frac{1 + N}{2N}\Big)^{\frac{N+1}{N-1}} 2^{\frac{2N}{1-N}} e_0^{\frac{2N}{N-1}}\Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}}\int_\Omega \Sigma(x)^{\frac{2}{1-N}} b(x)^{\frac{1+N}{1-N}}dx \le 0.\end{aligned} \tag{4.12}$$

We now restrict our consideration to an M phase composite. Let $\chi_r(x)$ be the indicator of phase r. The probability that a point x is in phase r is p_r so that the volume fraction of phase r is p_r. The probability that two points x and x' are in phases r and s, respectively, is $\langle\chi_r(x)\chi_s(x')\rangle = p_{rs}(x, x')$. The composite is assumed to be statistically homogeneous isotropic so that $p_{rs}(x,x') = p_{rs}(x-x') = p_{rs}(|x-x'|)$. The function $b(x)$ is constant in each phase, i.e.,

$$b(x) = \sum_{r=1}^{M} b_r \chi_r(x).$$

We also restrict the field $\beta(x)$ to the form

$$\beta(x) = \sum_{r=1}^{M} \beta^r \chi_r(x).$$

The expression (4.11) that we need to maximize has the form

$$\frac{1}{2}\int (M+N)_{ijkl}(x)\left(\sum_{r,s=1}^{M} \beta^r_{ij}\beta^s_{kl}p_{rs}(x) - \beta^r_{ij}\beta^s_{kl}p_r p_s\right)dx + \sum_{s=1}^{M} p_s\beta^s_{ij}\bar{\sigma}_{ij} - \sum_{s=1}^{M}\frac{\beta^s_{ij}\beta^s_{ij}}{4(b_s-b_0)}p_s.$$

Taking a variation with respect to β, we deduce that the best value of β satisfies the equation

$$\sum_{r,s=1}^{M} {\beta'}^s_{ij}\beta^r_{kl}\int (M+N)_{ijkl}(x)(p_{rs}(x) - p_r p_s)dx + \left(\sum_{s=1}^{M} p_s {\beta'}^s_{ij}\right)\bar{\sigma}_{ij} - \sum_{s=1}^{M}\frac{\beta^s_{ij}{\beta'}^s_{ij}}{2(b_s-b_0)}p_s = 0$$

for all ${\beta'}^s$ for $s = 1,\ldots,M$ so that

$$\sum_{r=1}^{M}\beta^r_{kl}\int (M+N)_{ijkl}(x)(p_{rs}(x) - p_r p_s)dx + p_s\bar{\sigma}_{ij} - \frac{\beta^s_{ij}p_s}{2(b_s-b_0)} = 0.$$

Restricting to a two-phase isotropic material, there is a function $h(x) = h(|x|)$ such that

$$p_{11}(x) - p_1p_1 = p_{22}(x) - p_2p_2 = -(p_{12}(x) - p_1p_2) = -(p_{21}(x) - p_1p_2) = p_1p_2h(r),$$

where $r = |x|$. Defining the fourth-order tensor

$$A_{ijkl} = \int (M+N)_{ijkl}(x)h(x)dx,$$

the equations for β then become

$$A_{ijkl}(\beta^s_{kl} - \langle\beta_{kl}\rangle) + \bar{\sigma}_{ij} - \frac{\beta^s_{ij}}{2(b_s-b_0)} = 0.$$

To determine the left-hand side of (4.12), we need to compute $\langle\beta\rangle$. In Hill's notation, $A = (0, 2\mu_A)$. The equations for β are written as

$$(0,2\mu_A)(\beta^s - \langle\beta\rangle) + \bar{\sigma} - \frac{\beta^s}{2(b_s-b_0)} = 0,$$

which gives

$$\beta^s = \frac{2(b_s-b_0)}{4\mu_A(b_s-b_0)-1}\Big((0,2\mu_A)\langle\beta\rangle - \bar{\sigma}\Big).$$

From this we get

$$\langle \beta \rangle = \frac{\sum_{s=1}^{2} p_s 2(b_s - b_0)/(4\mu_A(b_s - b_0) - 1)}{4\mu_A \sum_{s=1}^{2} p_s(b_s - b_0)/(4\mu_A(b_s - b_0) - 1) - 1} \bar{\sigma}.$$

Letting $b_0 = b_1$ and $b_2 = \gamma b_0$ where $\gamma > 1$,

$$\langle \beta \rangle = \frac{2p_2 b_0(\gamma - 1)}{1 - 4p_1\mu_A b_0(\gamma - 1)} \bar{\sigma}.$$

We now maximize the left-hand side of (4.12) with respect to b_0 and γ to get

$$\begin{aligned} &\sup_{b_0} \sup_{\gamma \geq 1} \Bigg\{ \frac{p_2 b_0(\gamma - 1)}{1 - 4p_1\mu_A b_0(\gamma - 1)} \bar{\sigma} : \bar{\sigma} + \frac{1}{4\mu} \bar{\sigma} : \bar{\sigma} + b_0 \bar{\sigma} : \bar{\sigma} - \frac{1}{2} \bar{\sigma} : \bar{\varepsilon} \\ &- \frac{1 - N}{2N} \Big(\frac{1 + N}{2N}\Big)^{\frac{N+1}{N-1}} \Big(\frac{2}{e_0}\Big)^{\frac{2N}{1-N}} \Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}} b_0^{\frac{1+N}{1-N}} \Big(\Sigma_1^{\frac{2}{1-N}} p_1 + \gamma^{\frac{1+N}{1-N}} \Sigma_2^{\frac{2}{1-N}} p_2\Big) \Bigg\} \leq 0. \end{aligned} \tag{4.13}$$

A bound for the modulus of the stress can then be found from

$$\begin{aligned} &\sup_{b_0} \sup_{\gamma \geq 1} \Bigg\{ \frac{p_2 b_0(\gamma - 1)}{1 - 4p_1\mu_A b_0(\gamma - 1)} \|\sigma\|^2 + \frac{1}{4\mu} \|\sigma\|^2 + b_0 \|\sigma\|^2 - \frac{1}{2} \|\sigma\| \|\bar{\varepsilon}\| \\ &- \frac{1 - N}{2N} \Big(\frac{1 + N}{2N}\Big)^{\frac{N+1}{N-1}} \Big(\frac{2}{e_0}\Big)^{\frac{2N}{1-N}} \Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}} b_0^{\frac{1+N}{1-N}} \Big(\Sigma_1^{\frac{2}{1-N}} p_1 + \gamma^{\frac{1+N}{1-N}} \Sigma_2^{\frac{2}{1-N}} p_2\Big) \Bigg\} \leq 0. \end{aligned} \tag{4.14}$$

In Appendix B, we show that when the function $h(r) = e^{-r/a}$,

$$\mu_A = \frac{3}{20b_0} \frac{1}{(l/a + (1 + 4\mu b_0)^{1/2})^2} + \frac{1}{10b_0} \frac{1}{(l/a + 1)^2} - \frac{1}{4b_0}.$$

The quantity a represents the correlation length scale for the microstructure. When $|x - x'| >> a$, $p_{rs}(x, x') \approx p_r(x)p_s(x')$. Since we use a linear comparison material whose strength reference is of the order $1/b_0$, deep in the plastic range, we expect that the inverse of the optimal b_0 is of the order of the secant modulus which is much less than μ, i.e., $\mu b_0 >> 1$. Therefore, only the last two terms in the expression of μ_A have significant contribution; μ_A is approximated as

$$\mu_A = \frac{1}{10b_0} \frac{1}{(l/a + 1)^2} - \frac{1}{4b_0}.$$

We then take the maximum of the sum of the terms involving b_0 in (4.13)

$$\begin{aligned} &\frac{5\gamma(l/a + 1)^2 - 2p_1(\gamma - 1)}{(p_2 + \gamma p_1)5(l/a + 1)^2 - 2p_1(\gamma - 1)} b_0 \bar{\sigma} : \bar{\sigma} \\ &- \frac{1 - N}{2N} \Big(\frac{1 + N}{2N}\Big)^{\frac{N+1}{N-1}} \Big(\frac{2}{e_0}\Big)^{\frac{2N}{1-N}} \Big(\frac{2}{3}\Big)^{\frac{1+N}{1-N}} b_0^{\frac{1+N}{1-N}} \Big(\Sigma_1^{\frac{2}{1-N}} p_1 + \gamma^{\frac{1+N}{1-N}} \Sigma_2^{\frac{2}{1-N}} p_2\Big) \end{aligned}$$

with respect to b_0. The approximated bound is

$$\Sigma^+ e_0 \Big(\frac{2}{3}\Big)^{-\frac{1+N}{2N}} (\bar{\sigma} : \bar{\sigma})^{\frac{1+N}{2N}} - \bar{\sigma} : (\bar{\varepsilon} - \frac{\bar{\sigma}}{2\mu}) \leq 0, \tag{4.15}$$

where

$$\Sigma^+ = \max_{\gamma\ge 1}\left\{\left(\frac{5\gamma(l/a+1)^2-2p_1(\gamma-1)}{(p_2+\gamma p_1)5(l/a+1)^2-2p_1(\gamma-1)}\right)^{\frac{1+N}{2N}}\left(p_1\Sigma_1^{\frac{2}{1-N}}+p_2\gamma^{\frac{1+N}{1-N}}\Sigma_2^{\frac{2}{1-N}}\right)^{\frac{N-1}{2N}}\right\}.$$

Using the Cauchy–Schwarz inequality, a bound for $\|\bar\sigma\|$ can be found from

$$\Sigma^+ e_0\Big(\frac{2}{3}\Big)^{-\frac{1+N}{2N}}\|\bar\sigma\|^{\frac{1+N}{N}}-\|\bar\sigma\|\|\bar\epsilon\|+\frac{1}{2\mu}\|\bar\sigma\|^2\le 0. \tag{4.16}$$

Regarding

$$\bar\varepsilon-\frac{\bar\sigma}{2\mu}=\bar\varepsilon^P,$$

and using the Cauchy–Schwarz inequality, we get the bound for the modulus of the stress

$$\|\sigma\|\le(\Sigma^+)^{-N}e_0^{-N}\Big(\frac{2}{3}\Big)^{\frac{1+N}{2}}\|\bar\varepsilon^P\|^N. \tag{4.17}$$

This bound can be obtained by differentiating the energy upper bound U^+ in (5.12) of [3] with respect to $\bar\varepsilon^P$ ($\bar\varepsilon^P$ here stands for ε^{P0} in [3]) when we disregard the elastic part; i.e., it equals

$$\frac{\partial U^+}{\partial\bar\varepsilon^P_{ij}}\frac{\partial U^+}{\partial\bar\varepsilon^P_{ij}}.$$

5. Numerical results.

We study the effect of the length scale l on the stress bounds found in the previous sections. By differentiating the Hashin–Shtrikman upper bound for the energy, in the case of macroscopic uniaxial response, Fleck and Willis [3] show that when the material length scale l increases, the stress increases. This is confirmed from our bound (4.14) for the magnitude $\|\bar\sigma\|$ of the stress. In Figure 5.1, we plot on the log-log axes the magnitude of the stress $\|\bar\sigma\|$ versus the magnitude of the strain $\|\bar\varepsilon\|$ for the example where $N = 0.3$, $p_1 = 0.3$, $p_2 = 0.7$, $\Sigma_1 = 0.03$, $\Sigma_2 = 0.01$, $e_0 = 1$, $\mu = 1$ for the values 0.01, 0.1, and 1 of l/a. It is clear that the stress bound increases with increasing l/a. An additional feature that manifests from this plot is that when the total strain magnitude $\|\bar\varepsilon\|$ exceeds 0.015, the plots asymptote straight lines; i.e., the stress-strain relation follows a power law. This is consistent with (4.17), where we made the approximation (4.15) of the stress bound. (Note that deep in the plastic range, the magnitude of $\bar\varepsilon^P$ approximates $\|\bar\varepsilon\|$.)

Similar features are shown in Figure 5.2 for the same parameters except that $N = 0.1$. In the range plotted, since N is now smaller, we see that for small $\|\bar\varepsilon\|$, the bounds for different values of l/a are almost the same. The elastic strain dominates in this range so the material scale l has no effect.

Figure 5.3 shows that the approximated formula (4.16) gives about the same bound for the magnitude of the stress when the magnitude of the strain exceeds about 0.015. The dashed line is the stress versus strain curve found from the (4.16) where the solid line is the exact bound (4.14). When the total stress is smaller than 0.01, (4.16) may not adequately represent (4.14). The parameters we use are $N = 0.3, p_1 = 0.3, p_2 = 0.7, \Sigma_1 = 0.03, \Sigma_2 = 0.01, l/a = 0.1, e_0 = 1, \mu = 1$.

Keeping the same parameters except that $N = 0.1$, Figure 5.4 shows the exact Hashin–Shtrikman bound (4.14) and the approximated bound (4.16) for the magnitude of the stress in the log-log axis. We see that the two bounds are almost the same for small strain. This is because of the dominance of the elastic strain.

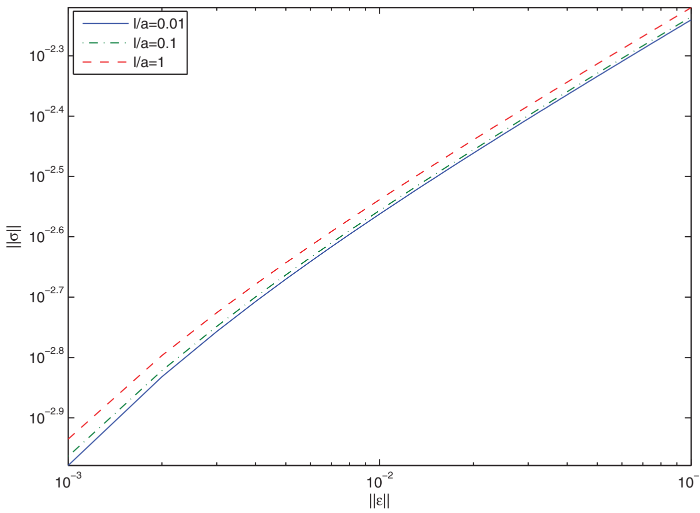


FIG. 5.1. *The stress bound increases with l/a.*

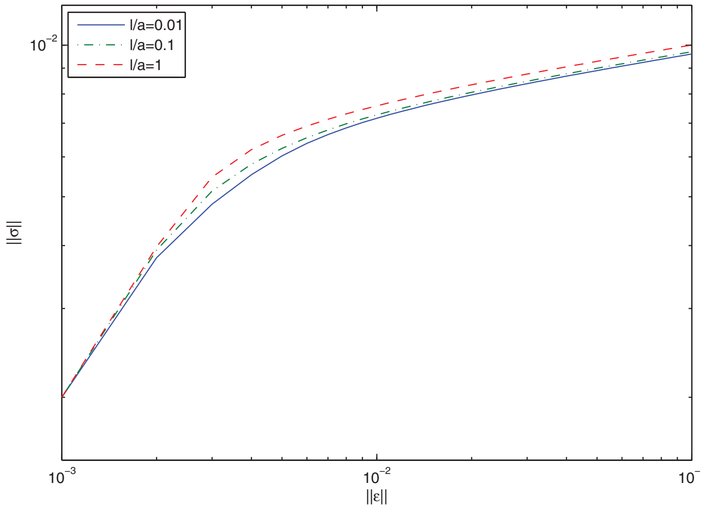


FIG. 5.2. *The stress versus the strain for the exact Hashin–Shtrikman bounds ($N = 0.1$).*

Given the boundary strain $\bar{\varepsilon}$, (3.3) and (4.13) give a region in the stress space where the effective stress tensor should lie for all the microstructures. In Figure 5.5, we plot the area restricted by the bounds (3.3) and (4.13) for the two-dimensional stress space $\bar{\sigma} = (\sigma_1, \sigma_2)$. The two-dimensional strain $\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2)$ is $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 0.0005$. This corresponds to the small stress magnitude area in Figure 5.3. Other parameters are $N = 0.3$, $l/a = 0.1$, $p_1 = 0.3$, $p_2 = 0.7$, $\Sigma_1 = 0.03$, $\Sigma_2 = 0.01$, $e_0 = 1$, $\mu = 1$.

Figure 5.6 shows the area of the stress for $\varepsilon_1 = 0.08$ and $\varepsilon_2 = 0.04$. The stress magnitude is in the larger end in Figure 5.3. Other parameters are kept the same. The area bounded by (3.3) seems to blow up as the magnitude of $\bar{\sigma}$ is now much

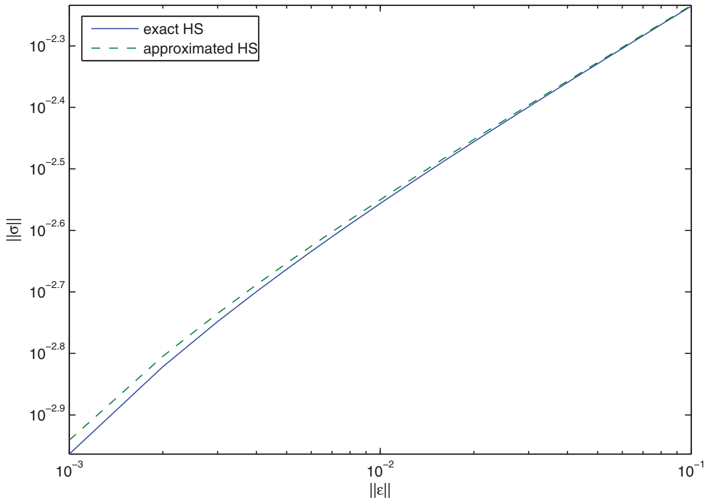


FIG. 5.3. *Stress versus strain for the exact and approximated Hashin–Shtrikman-type bounds.*

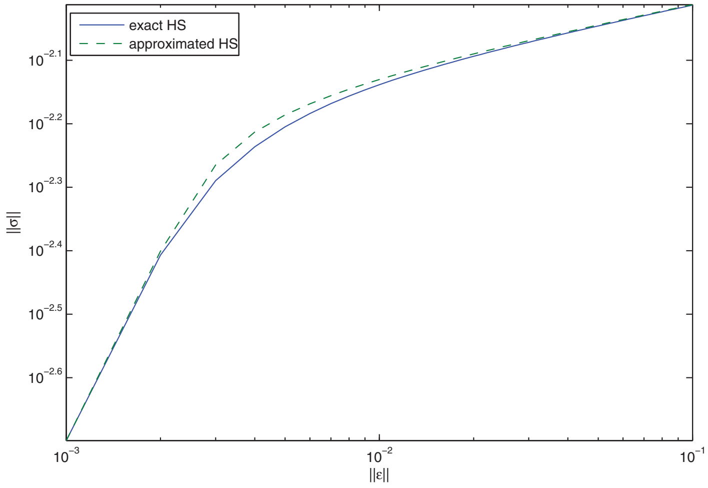


FIG. 5.4. *Stress versus strain for the exact and approximated Hashin–Shtrikman bounds ($N = 0.1$).*

smaller than the magnitude of $\bar{\varepsilon}$. The bound (3.3) now behaves like the linear bound

$$(\bar{\varepsilon} : \bar{\varepsilon})^{\frac{1+N}{2}} \left(\frac{2}{3}\right)^{\frac{N+1}{2}} e_0^{-N} \langle \Sigma^{-1/N} \rangle^{-N} \leq \bar{\sigma} : \bar{\varepsilon};$$

this is shown in Figure 5.7.

The bound (3.3) is insensitive to l/a. As shown in Figure 5.1, the bound for $\|\sigma\|$ increases as l/a increases. We predict that given a boundary strain $\bar{\varepsilon}$, the area in the stress space that bounds $\bar{\sigma}$ gets larger with increasing l/a. This is shown in Figure 5.8 for $\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2)$ where $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 0.0005$. Other parameters are as for Figure 5.6. The area for $l/a = 0.1$ is strictly contained inside the area for $l/a = 1$.

For the same parameters, Figure 5.9 presents the bound for the stress $\bar{\sigma} = (\sigma_1, \sigma_2)$ given that $\bar{\varepsilon} = (0.001, 0.0005)$. The dashed line shows the bound given by the ap-

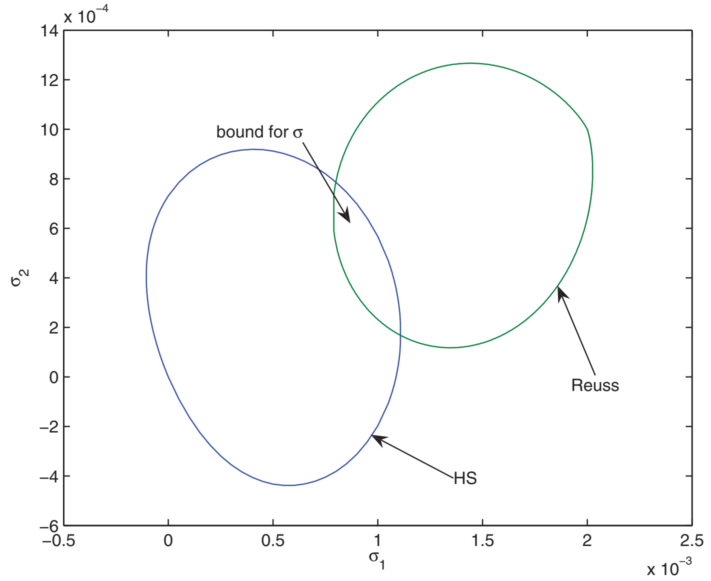


FIG. 5.5. *Bound for the stress when* $\bar{\varepsilon} = (0.001, 0.0005)$.

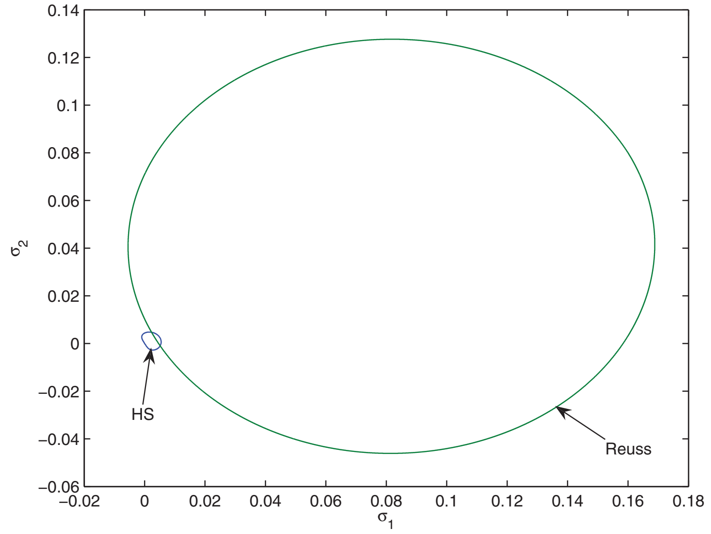


FIG. 5.6. *The stress bounds* (3.3) *and* (4.13) *in the two-dimensional space for* $\bar{\varepsilon} = (0.08, 0.04)$.

proximated stress bound (4.15). It is clear that in the small stress magnitude range, (4.15) does not give a good approximation to the exact bound. However, for large stress magnitude, (4.15) gives an excellent approximation. This is shown in Figure 5.10 for $\bar{\varepsilon} = (0.08, 0.04)$.

6. Conclusions. We established bounds for the average stress in terms of the average strain for an elasto-plastic composite whose components follow the strain gradient theory of Fleck and Willis [3]. We confirmed rigorously the prediction by [3] on the dependence of the average stress on the internal length scale, made by

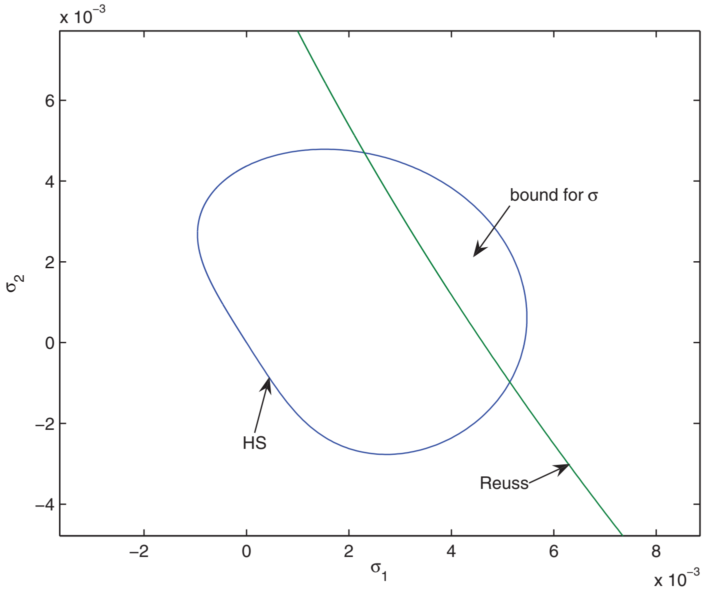


Fig. 5.7. *Bound for the stress when* $\bar{\varepsilon} = (0.08, 0.04)$.

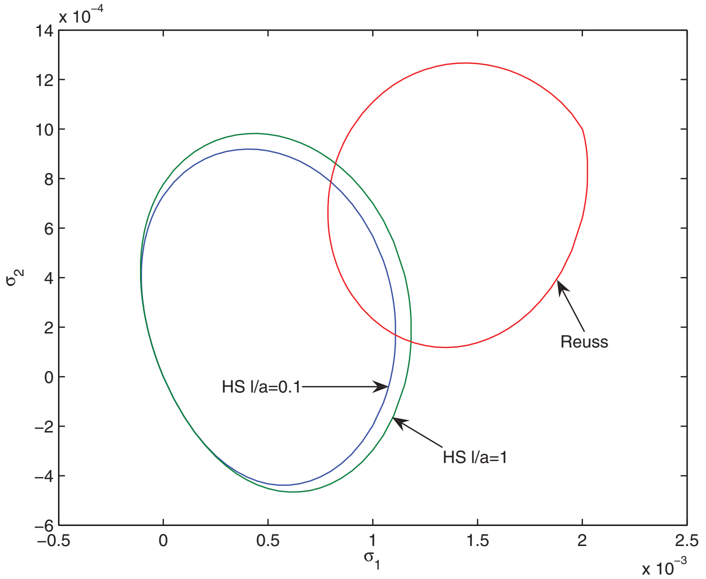


Fig. 5.8. *Bound for the the stress increases when* l/a *increases.*

formally differentiating their energy bounds. Our results provide stress bounds for all the range of the plastic strain. Deep in the plastic regime, they coincide with the formal approximations by Fleck and Willis [3]. Given an average strain, the bounds obtained here provide a specific region in the stress space in which the average stress must lie, which is not available from the energy bounds. Our method is a combination of the known bounding approaches for composite materials. It is a modification of the elegant techniques for bounding the constitutive relations by Milton and Serkov

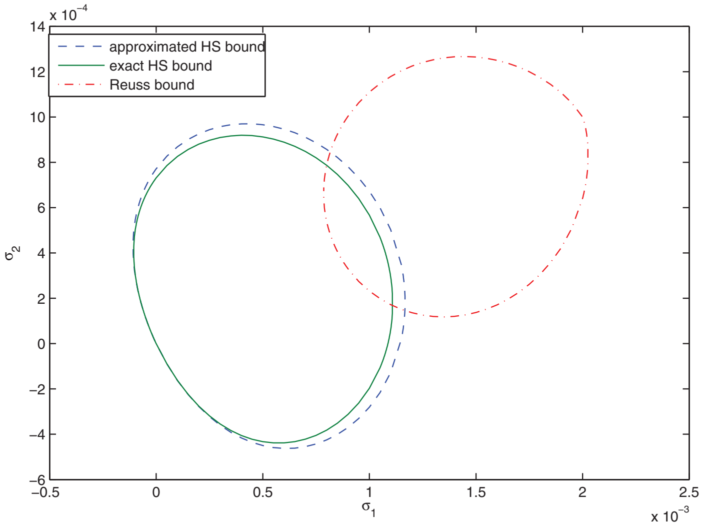


FIG. 5.9. *Exact and approximated bounds for* $\bar{\varepsilon} = (0.001, 0.0005)$.

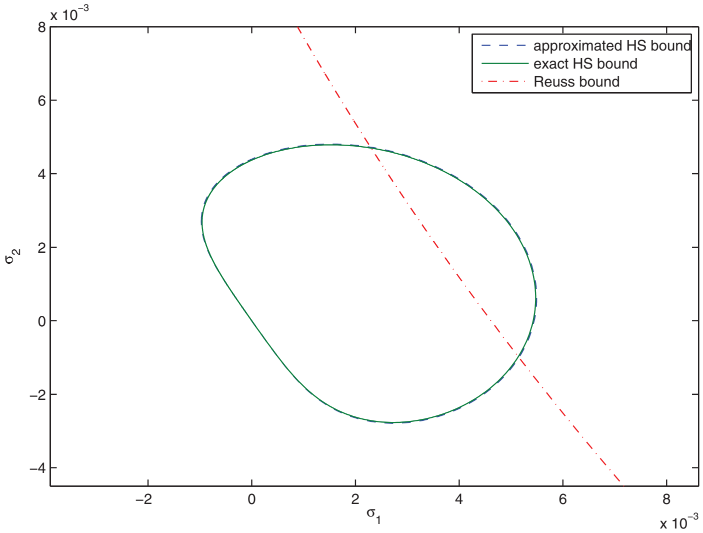


FIG. 5.10. *Exact and approximated bounds for* $\bar{\varepsilon} = (0.08, 0.04)$.

[8] and Talbot and Willis [19]. It also follows the ideas developed for the "nonlinear Hashin–Shtrikman" bounds by Talbot and Willis [18], and Ponte-Castañeda [12].

Appendix A. In this appendix, we solve the equations (4.7) and (4.9). Let $G_{ip}(x, x')$ be the Green function satisfying

$$\frac{\partial^2 G_{ip}(x,x')}{\partial x_j \partial x_j} + \delta_{ip}\delta(x - x') = P_{,i},$$

subjecting to the incompressible condition $\partial G_{ip}(x,x')/\partial x_i = 0$ with the homogeneous boundary condition. The solution v for (4.7) can be written as

$$v_i(x') = \bar{e}_{ij}x_j + \int_\Omega \frac{\partial G_{ki}(x,x')}{\partial x_l}(2\beta_{kl}(x) + 4b_0 s_{kl}(x))dx.$$

Since

$$\int_\Omega \frac{\partial G_{ki}(x,x')}{\partial x_l}dx = 0,$$

we can write v as

$$v_i(x') = \bar{e}_{ij}x_j + \int_\Omega \frac{\partial G_{ki}(x,x')}{\partial x_l}(2\beta_{kl}(x) + 4b_0 s_{kl}(x) - 2\langle\beta_{kl}\rangle)dx.$$

Let

$$e_{ij}(x') = \frac{1}{2}(v_{i,j}(x') + v_{j,i}(x')).$$

Differentiating both sides and taking the symmetrization, we get

$$e_{ij}(x') = \bar{e}_{ij} + \int_\Omega \Gamma_{ijkl}(x,x')(2\beta_{kl}(x) + 4b_0 s_{kl}(x) - 2\langle\beta_{kl}\rangle)dx \tag{A.1}$$

where the fourth-order tensor Γ is defined as

$$\Gamma_{ijkl}(x,x') = \frac{\partial^2 G_{ki}(x,x')}{\partial x_l \partial x'_j} \quad \text{symm } (i,j)(k,l).$$

Assuming that the correlation length is small, following [3], we can substitute Γ by its infinite body form, which depends only on $x - x'$; i.e., $\Gamma(x,x') = \Gamma(x-x')$. Then from (4.6) and (4.8), we have

$$\begin{aligned} &\beta_{ij} + \frac{1}{2\mu}\sigma_{ij} + 2b_0(\sigma_{ij} + s_{ij}) = \langle\beta_{ij}\rangle + \frac{\bar{\sigma}_{ij}}{2\mu} + 2b_0\bar{\sigma}_{ij} \\ &\quad + \int \Gamma_{ijkl}(x-x')(2\beta_{kl}(x) + 4b_0 s_{kl}(x) - 2\langle\beta_{kl}\rangle)dx. \end{aligned} \tag{A.2}$$

Note that the integral now can be approximated by an integral over the infinite domain.

To solve (4.9), we consider the function $g(x,x')$ that satisfies the equation

$$\nabla_x^2 g(x,x') - \frac{1}{l^2}g(x,x') + \delta(x-x') = 0,$$

with the Neumann boundary condition. The solution of (4.9) can be writen as

$$s_{ij}(x') + \sigma_{ij}(x') + \frac{\beta_{ij}(x')}{2b_0} = \frac{1}{l^2}\int_\Omega g(x,x')\Big(\sigma_{ij}(x) + \frac{\beta_{ij}(x)}{2b_0}\Big)dx.$$

Since $(1/l^2)\int_\Omega g(x,x')dx = 1$, we can write this as

$$s_{ij}(x') + \sigma_{ij}(x') + \frac{\beta_{ij}(x')}{2b_0} = \frac{1}{l^2}\int_\Omega g(x,x')\Big(\sigma_{ij}(x) + \frac{\beta_{ij}(x)}{2b_0} - \bar{\sigma}_{ij} - \frac{\langle\beta_{ij}\rangle}{2b_0}\Big) + \bar{\sigma}_{ij} + \frac{\langle\beta_{ij}\rangle}{2b_0}.$$

In the asymptotic limit, we can substitute $g(x, x')$ by its infinite body form $g(x - x')$. The equation then becomes

$$\begin{aligned} & s_{ij}(x') + \sigma_{ij}(x') + \frac{\beta_{ij}(x')}{2b_0} \\ & = \frac{1}{l^2} \int_\Omega g(x - x') \left(\sigma_{ij}(x) + \frac{\beta_{ij}(x)}{2b_0} - \bar{\sigma}_{ij} - \frac{\langle \beta_{ij} \rangle}{2b_0} \right) + \bar{\sigma}_{ij} + \frac{\langle \beta_{ij} \rangle}{2b_0}. \end{aligned} \tag{A.3}$$

We now need to solve (A.2) and (A.3) for σ and s. Taking the Fourier transform

$$\hat{f}(\xi) = \int e^{i\xi.x} f(x) dx,$$

we then have

$$\begin{aligned} \hat{\beta}_{ij} + \frac{\hat{\sigma}_{ij}}{2\mu} + 2b_0(\hat{\sigma}_{ij} + \hat{s}_{ij}) = \widehat{\langle \beta \rangle}_{ij} + \frac{\hat{\bar{\sigma}}_{ij}}{2\mu} + 2b_0 \hat{\bar{\sigma}}_{ij} \\ + \hat{\Gamma}_{ijkl}(2\hat{\beta}_{kl} + 4b_0 \hat{s}_{kl} - 2\widehat{\langle \beta \rangle}_{kl}), \end{aligned}$$

and

$$\hat{\sigma}_{ij} + \hat{s}_{ij} + \frac{\hat{\beta}_{ij}}{2b_0} = \frac{1}{l^2} \hat{g} \left(\hat{\sigma}_{ij} + \frac{\hat{\beta}_{ij}}{2b_0} - \hat{\bar{\sigma}}_{ij} - \frac{\widehat{\langle \beta \rangle}_{ij}}{2b_0} \right) + \hat{\bar{\sigma}}_{ij} + \frac{\widehat{\langle \beta \rangle}_{ij}}{2b_0}.$$

Rearranging these equations we have

$$\hat{\sigma} - \hat{\bar{\sigma}} = \frac{2\mu(2\hat{\Gamma} - I)}{1 + 4b_0\mu}(\hat{\beta} - \widehat{\langle \beta \rangle}) + \frac{4\mu b_0(2\hat{\Gamma} - I)}{1 + 4b_0\mu} \hat{s},$$

and

$$\hat{\sigma} - \hat{\bar{\sigma}} = \frac{-1}{2b_0}(\hat{\beta} - \widehat{\langle \beta \rangle}) - \frac{l^2}{l^2 - \hat{g}} \hat{s},$$

where

$$I_{ijkl} = \frac{1}{2} \left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} - \frac{2}{3}\delta_{ij}\delta_{kl} \right)$$

is the incompressible identity tensor. From this

$$\left(\frac{2\mu(2\hat{\Gamma} - I)}{1 + 4b_0\mu} + \frac{I}{2b_0} \right)(\hat{\beta} - \widehat{\langle \beta \rangle}) + \left(\frac{4\mu b_0(2\hat{\Gamma} - I)}{1 + 4\mu b_0} + \frac{l^2 I}{l^2 - \hat{g}} \right) \hat{s} = 0. \tag{A.4}$$

On using $\hat{\Gamma}(2\hat{\Gamma} - I) = 0$,

$$\hat{\Gamma}\hat{s} = -\frac{l^2 - \hat{g}}{2b_0 l^2} \hat{\Gamma}(\hat{\beta} - \widehat{\langle \beta \rangle}).$$

Substituting this into (A.4), we get

$$\frac{8\mu b_0 \hat{g} \hat{\Gamma} + l^2 I}{2b_0 l^2 (1 + 4\mu b_0)}(\hat{\beta} - \widehat{\langle \beta \rangle}) + \frac{l^2 + 4\mu b_0 \hat{g}}{(l^2 - \hat{g})(1 + 4\mu b_0)} \hat{s} = 0.$$

From this we find

$$\hat{s} = \frac{-(8\mu b_0 \hat{g}\hat{\Gamma} + l^2 I)(l^2 - \hat{g})}{2b_0 l^2 (l^2 + 4\mu b_0 \hat{g})}(\hat{\beta} - \widehat{\langle\beta\rangle}), \tag{A.5}$$

and

$$\hat{\sigma} - \hat{\bar{\sigma}} = \frac{2\mu \hat{g}(2\hat{\Gamma} - I)}{l^2 + 4\mu b_0 \hat{g}}(\hat{\beta} - \widehat{\langle\beta\rangle}), \tag{A.6}$$

which we denote as

$$\hat{s} = \hat{N}(\hat{\beta} - \widehat{\langle\beta\rangle}),$$

and

$$\hat{\sigma} = \hat{M}(\hat{\beta} - \widehat{\langle\beta\rangle}).$$

The values of $\hat{\Gamma}$ and $\hat{g}$ are

$$\hat{g}(\xi) = \frac{l^2}{1 + l^2|\xi|^2}, \quad \hat{\Gamma}_{ijkl}(\xi) = \frac{\delta_{ik}\xi_j\xi_l}{|\xi|^2} - \frac{\xi_i\xi_j\xi_k\xi_l}{|\xi|^4}.$$

Appendix B. We now compute the value of μ_A. We have

$$10\mu_A = A_{ijij} = \int (M+N)_{ijij}(x)h(x)dx.$$

From (A.5) and (A.6),

$$(\hat{M} + \hat{N})_{ijij} = \frac{2\mu \hat{g}^2 (2\hat{\Gamma}_{ijij} - I_{ijij})}{l^2(l^2 + 4\mu b_0 \hat{g})} + \frac{\hat{g} I_{ijij}}{2b_0 l^2} - \frac{I_{ijij}}{2b_0}.$$

Since $\hat{\Gamma}_{ijij} = 1$ and $I_{ijij} = 5$,

$$(\hat{M} + \hat{N})_{ijij} = \frac{3}{2b_0}\frac{1}{l^2|\xi|^2 + 1 + 4\mu b_0} + \frac{1}{b_0(l^2|\xi|^2 + 1)} - \frac{5}{2b_0}.$$

We then apply the same procedure as in [3].

Acknowledgment. The author is grateful to John R. Willis for fruitful discussions.

REFERENCES

[1] K. E. Aifantis and J. R. Willis, *The role of interfaces in enhancing the yield strength of composites and polycrystals*, J. Mech. Phys. Solids, 53 (2005), pp. 1047–1070.

[2] N. A. Fleck and J. W. Hutchinson, *A reformulation of strain gradient plasticity*, J. Mech. Phys. Solids, 49 (2001), pp. 2245–2272.

[3] N. A. Fleck and J. R. Willis, *Bounds and estimates for the effect of strain gradients upon the effective plastic properties of an isotropic two phase composite*, J. Mech. Phys. Solids, 52 (2004), pp. 1855–1888.

[4] P. Gudmundson, *A unified treatment of strain gradient plasticity*, J. Mech. Phys. Solids, 52 (2004), pp. 1379–1406.

[5] R. B. Kohn and T. D. Little, *Some model problems of polycrystal plasticity with deficient basic crystals*, SIAM J. Appl. Math., 59 (1998), pp. 172–197.

[6] K. A. Lurie and A. V. Cherkaev, *Accurate estimates of the conductivity of mixtures formed of two materials in a given proportion (two dimensional problem)*, Dokl. Akad. Nauk., 264 (1982), pp. 1128–1130.

[7] K. A. Lurie and A. V. Cherkaev, *Exact estimates of conductivity of composites formed by two isotropically conducting medium taken in prescribed proportion*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 71–87.

[8] G. W. Milton and S. K. Serkov, *Bounding the current in nonlinear conducting composites*, J. Mech. Phys. Solids, 48 (2000), pp. 1295–1324.

[9] F. Murat and L. Tartar, *Calcul des variations et homogénisation. Les méthodes de l'homogénéisation: Théorie et applications en physique*, Coll. de la Dir des Études et Recherches de Électricité de France, Eyrolles, Paris, 1985, pp. 319-370. (Translated in *Topics in the Mathematical Modelling of Composite Materials*, A. Cherkaev and R. Kohn, eds., Progress in Nonlinear Differential Equations and Their Applications, Basel: Birkhauser, 31, pp. 139–173).

[10] V. Nesi, V. P. Smyshlyaev, and J. W. Willis, *Improved bounds for the yield stress of a model polycrystalline material*, J. Mech. Phys. Solids, 48 (2000), pp. 1799–1825.

[11] M. Peigney, *A pattern-based method for bounding the effective response of a nonlinear composite*, J. Mech. Phys. Solids, 53 (2005), pp. 923–948.

[12] P. Ponte Castañeda, *New variational principles in plasticity and their application to composite*, J. Mech. Phys. Solids, 40 (1992), pp. 1757–1788.

[13] P. Ponte Castañeda and G. De Botton, *On the homogenized yield strength of two phase composites*, Proc. Roy. Soc. A, 438 (1992), pp. 419–431.

[14] V. P. Smyshlyaev and N. A. Fleck, *Bounds and estimates for linear composites with strain gradient effects*, J. Mech. Phys. Solids, 42 (1994), pp. 1851–1882.

[15] V. P. Smyshlyaev and N. A. Fleck, *Bounds and estimates for the overall plastic behaviour of composites with strain gradient effects*, Proc. Roy. Soc. A, 451 (1995), pp. 795–810.

[16] V. P. Smyshlyaev and N. A. Fleck, *The role of the strain gradients in the grain size effect for polycrystals*, J. Mech. Phys. Solids, 42 (1996), pp. 465–495.

[17] P. M. Suquet, *Overall potentials and extremal surfaces of power law or ideally plastic composites*, J. Mech. Phys. Solids, 41 (1993), pp. 981–1002.

[18] D. R. S. Talbot and J. R. Willis, *Variational principles for inhomogeneous nonlinear media*, IMA J. Appl. Math., 35 (1985), pp. 39–54.

[19] D. R. S. Talbot and J. R. Willis, *Bounds for the effective constitutive relation of a nonlinear composite*, Proc. R. Soc. Lond. A, 460 (2004), pp. 2705–2723.

[20] L. Tartar, *Estimation de coefficients homogénéisés*, in Computer Methods in Applied Sciences and Engineering, Lecture Notes in Mathematics, Springer Verlag, Berlin, 704 (1977), pp. 136–212. (Translated in *Topics in the Mathematical Modelling of Composite Materials*, A. Cherkaev and R. Kohn, eds., Progress in Nonlinear Differential Equations and Their Applications, Basel: Birkhauser, 31 (1997), pp. 139–173.)

[21] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, R. J. Knops, ed., Heriot Watt Symposium. Research Notes in Mathematics, Pitman, Boston, vol. IV (1979), pp. 136–212.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1618–1643

EFFECTIVE EQUATIONS FOR LOCALIZATION AND SHEAR BAND FORMATION*

THEODOROS KATSAOUNIS[†] AND ATHANASIOS E. TZAVARAS[‡]

Abstract. We develop a quantitative criterion determining the onset of localization and shear band formation at high strain-rate deformations of metals. We introduce an asymptotic procedure motivated by the theory of relaxation and the Chapman–Enskog expansion and derive an effective equation for the evolution of the strain rate, consisting of a second order nonlinear diffusion regularized by fourth order effects and with parameters determined by the degree of thermal softening, strain hardening, and strain-rate sensitivity. The nonlinear diffusion equation changes type across a threshold in the parameter space from forward parabolic to backward parabolic, what highlights the stable and unstable parameter regimes. The fourth order effects play a regularizing role in the unstable region of the parameter range.

Key words. shear band, localization, thermoviscoplasticity, Chapman–Enskog expansion

AMS subject classifications. 74C20, 74H40, 35K65, 35Q72

DOI. 10.1137/080727919

1. Introduction.

One striking instance of material instability is observed in the course of deformations of metals at high strain rates. It appears as an instability in shear and leads to regions of intensely concentrated shear strain, called shear bands. Since shear bands are often precursors to rupture, their study has attracted attention in the mechanics literature (e.g., [1, 7, 8, 13, 14, 17, 22, 24, 25]).

In experimental investigations of high strain-rate deformations of steels, observations of shear bands are typically associated with strain softening response—past a critical strain—of the measured stress-strain curve [8]. It was recognized by Zener and Hollomon [27] that the effect of the deformation speed is twofold: First, an increase in the deformation speed changes the deformation conditions from isothermal to nearly adiabatic. Second, strain rate has an effect per se and needs to be included in the constitutive modeling.

Under isothermal conditions, metals, in general, strain harden and exhibit a stable response. As the deformation speed increases, the heat produced by the plastic work causes an increase in the temperature. For certain metals, the tendency for thermal softening may outweigh the tendency for strain hardening and deliver net softening. A destabilizing feedback mechanism is then induced, which operates as follows (see [8]): Nonuniformities in the strain rate result in nonuniform heating. Since the material is softer at the hotter spots and harder at the colder spots, if heat diffusion is too weak to equalize the temperatures, the initial nonuniformities in the strain rate are, in turn, amplified. This mechanism tends to localize the total deformation into narrow regions. On the other hand, there is opposition to this process by "viscous effects"

*Received by the editors June 20, 2008; accepted for publication (in revised form) December 15, 2008; published electronically March 18, 2009. This work was partially supported by National Science Foundation Grants DMS-0555272 and DMS-0757227 and the program "Pythagoras" of the Greek Secretariat of Research.

http://www.siam.org/journals/siap/69-6/72791.html

[†]Department of Applied Mathematics, University of Crete and Institute for Applied and Computational Mathematics, FORTH, Heraklion, Greece (thodoros@tem.uoc.gr).

[‡]Department of Applied Mathematics, University of Crete and Institute for Applied and Computational Mathematics, FORTH, Heraklion, Greece. Current address: Department of Mathematics, University of Maryland, College Park, MD 20742 (tzavaras@math.umd.edu).

induced by strain-rate sensitivity. The outcome of the competition depends mainly on the relative weights of thermal softening, strain hardening, and strain-rate sensitivity, as well as the loading circumstances.

This qualitative scenario is widely accepted as the mechanism of shear band formation. However, despite several attempts, a quantitative explanation of the phenomenon of shear bands is presently lacking. Moreover, the above picture is somewhat imprecise in terms of what determines (or rules out) the onset of localization. It is this aspect of the problem that we attempt to address in the present work. We use the model

$$\begin{aligned} v_t &= \frac{1}{r}\,\sigma_x, \\ \theta_t &= \kappa\theta_{xx} + \sigma\gamma_t, \\ \gamma_t &= v_x, \end{aligned} \tag{1.1}$$

where r, κ are nondimensional constants and the stress is given by an empirical power law in the normalized form

$$\sigma = \theta^{-\alpha}\gamma^m\gamma_t^n, \tag{1.2}$$

appropriate for the flow rule of a viscoplastic material exhibiting thermal softening, strain hardening, and strain-rate sensitivity. The model and its relevance to the problem of shear band formation is explained in section 2.

There is extensive literature on the problem, including experimental [7, 14], mechanics and linearized analysis (e.g. [8, 1, 13, 17, 24, 25] and references therein), numerical [26, 11], as well as nonlinear analysis [9, 18, 19, 20, 4, 3] and asymptotic analysis studies [10, 12, 25]. With regard to the analysis of the shear band formation process, analytical results account for either the case where the forcing is effected by a boundary force [20, 22] causing a shear band at the boundary or in situations where the initial data involve a localization in shear (or in the temperature) and the subsequent evolution leads to an intensification process to a fully developed band [3, 23]. It is indicated by numerical evidence in [24] and the analysis in [20, 3] that a collapse of the stress-diffusion mechanisms is associated with the development of the bands. There is a class of special solutions to (1.1) describing uniform shearing

$$\begin{aligned} v_s &= x, \\ \gamma_s &= t + \gamma_0, \\ \theta_s &= \left[\theta_0^{1+\alpha} + \frac{1+\alpha}{m+1}\Big[(t+\gamma_0)^{m+1} - \gamma_0^{m+1}\Big]\right]^{\frac{1}{1+\alpha}}, \\ \sigma_s &= \theta_s^{-\alpha}(t)(t+\gamma_0)^m, \end{aligned} \tag{1.3}$$

and much of the analysis on (1.1) has centered on the issue of their stability. The form of (1.3) suggests the change of variables (3.3) that transforms the stability problem into the study of the asymptotic behavior for the reaction–diffusion-type system (3.4); see section 3. In the special case of a fluid with temperature-dependent viscosity ($m = 0$) the kinematic equation decouples from the remaining equations, and the problem reduces to the study of the simplified model (4.1)–(4.2). This simpler system is indeed the one that has been analyzed in most detail both analytically [3, 9, 18] but also in numerical investigations [26, 11]. Its rescaled variant (4.4) admits invariant rectangles in the parameter range $q = -\alpha + n > 0$ but misses this property in the range $q = -\alpha + n < 0$. It is this dichotomy that provides a quantitative threshold to stability, as shown in section 4: In the parameter range $q > 0$ the invariant rectangles yield asymptotic stability of the uniform shearing solution; cf. Theorem 4.1. By contrast, in the complementary region $q < 0$ moderate perturbations of the uniform solutions can lead to instability and formation of shear bands; cf. Theorem 4.2.

The analysis on invariant domains of section 4 suggests a connection of the present problem with the theory of relaxation systems (e.g., [5]) that turns out to be instrumental for understanding the onset of localization. This connection is studied in detail in section 5 and motivates the derivation of an effective equation for the onset of localization in section 6. We outline the result in the following: Let T be a parameter describing a time-scale, and consider a change of variables of the form

$$\theta(x,t) = (t+1)^{\frac{m+1}{\alpha+1}}\Theta^T\left(x, \frac{s(t)}{T}\right), \quad v_x(x,t) = V_x^T\left(x, \frac{s(t)}{T}\right), \tag{1.4}$$

where $s(t)$ is an appropriate rescaling of time (in fact, see (6.1) for the full transformation). The new functions $(U^T, \Theta^T, \Gamma^T, \Sigma^T)$, with $U^T = V_x^T$ satisfy the system (6.3). It is clear that if $(U^T, \Theta^T, \Gamma^T, \Sigma^T)$ stabilizes as $T \to \infty$, then its limiting profile will describe the asymptotic form of (v_x, θ, γ) as $t \to \infty$. This reduces the problem of studying the asymptotic behavior into the problem of identifying the large T behavior of (6.3), which lies within the realm of relaxation theory. Using a technique analogous to the Chapman–Enskog expansion (e.g., [5]), we show in section 6 that $U^T = V_x^T$ satisfies for $T >> 1$ and $r = O(T)$ the effective equation

$$\partial_s U = \partial_{xx}\left(c\,U^p + \frac{\lambda c^2}{T}(\beta s + 1)U^{p-1}\partial_{xx}U^p\right), \tag{1.5}$$

within order $O(\frac{1}{T^2})$ and with parameters $p = \frac{q}{1+\alpha} = \frac{-\alpha+m+n}{1+\alpha}$, $\beta = \frac{m+1}{1+\alpha}$, $c = \beta^{\frac{\alpha}{1+\alpha}}$, and the coefficient of the fourth order term

$$\lambda = \frac{\alpha(1+m+n) - m(m+1)}{(m+1)(1+\alpha)}.$$

We note that (1.5) changes type from forward parabolic when $q = -\alpha + m + n > 0$ to backward parabolic when $q = -\alpha + m + n < 0$, what captures the parameter regime associated with the onset of localization. We also note that in the region of instability $q < 0$, the coefficient $\lambda > 0$, and the fourth order term has a regularizing effect. Numerical comparisons between the effective equation (1.5) and the system (6.3) are performed in section 6 and indicate good agreement between the effective and the actual problem.

2. The nature of shear band formation.

2.1. Description of the model.

As shear bands appear and propagate as one-dimensional structures (up to interaction times), most investigations have focused on one-dimensional, simple shearing deformations. In a Cartesian coordinate system an infinite plate, located between the planes $x = 0$ and $x = d$, is subjected to simple shear. The thermomechanical process is described (upon neglecting the normal stresses) by the list of variables: Velocity in the shearing direction $v(x,t)$, shear strain $\gamma(x,t)$, temperature $\theta(x,t)$, heat flux $Q(x,t)$, and shear stress $\sigma(x,t)$. They are connected through the balance of linear momentum

$$\rho v_t = \sigma_x, \tag{2.1}$$

the kinematic compatibility relation

$$\gamma_t = v_x, \tag{2.2}$$

and the balance of energy equation

$$c\rho\theta_t = Q_x + \beta\sigma\gamma_t, \tag{2.3}$$

where ρ is the reference density, c is the specific heat, and β is the portion of plastic work converted to heat. The upper plate is subjected to a prescribed constant velocity V, while the lower plate is held at rest: $v(0,t) = 0$, $v(d,t) = V$. It is further assumed that the plates are thermally insulated: $\theta_x(0,t) = 0$, $\theta_x(d,t) = 0$. Thermal insulation is appropriate for the analysis of shear band formation, since heat transfer at positions distant from the bands via radiation is negligible at sufficiently short loading times.

For the heat flux we will use either the adiabatic assumption $Q = 0$ or a Fourier law $Q = k\theta_x$, with the thermal diffusivity parameter k. Imposing adiabatic conditions projects the belief that, at high strain rates, heat diffusion operates at a slower timescale than the one required for the development of a shear band. It appears a plausible assumption for the shear band initiation process but not necessarily for the evolution of a developed band, due to the high temperature differences involved.

For the shear stress we set

$$\sigma = f(\theta, \gamma, \gamma_t), \tag{2.4}$$

where f is a smooth function, with $f(\theta,\gamma,0) = 0$ and $f_p(\theta,\gamma,p) > 0$ for $p \neq 0$. In terms of classification, the resulting model belongs to the framework of one-dimensional thermoviscoelasticity and is compatible with the requirements imposed by the Clausius–Duhem inequality.

It is instructive to interpret (2.4) in the context of a constitutive theory for thermal elastic-viscoplastic materials. In this context, it is assumed that the shear strain γ is decomposed, additively, into elastic and plastic components: $\gamma = \gamma^e + \gamma^p$. The elastic component γ^e satisfies linear elasticity with shear modulus G_e, that is, $\gamma^e = \frac{1}{G_e}\sigma$. The evolution of the plastic component is dictated by a plastic flow rule:

$$\gamma_t^p = g(\theta, \gamma^p, \sigma) \quad \text{or} \quad \sigma = f(\theta, \gamma^p, \gamma_t^p), \tag{2.5}$$

where g is an increasing function in the variable σ and $f(\theta,\gamma,\cdot)$ is the inverse function of $g(\theta,\gamma,\cdot)$. In summary,

$$\begin{aligned} \frac{1}{G_e}\sigma + \gamma^p &= \gamma, \\ \frac{1}{G_e}\sigma_t + g(\theta, \gamma^p, \sigma) &= v_x. \end{aligned} \tag{2.6}$$

Note that (2.4) can be obtained from the constitutive theory (2.6) in the limit as the elastic shear modulus $G_e \to \infty$. Accordingly, γ should then be interpreted as the plastic strain and (2.4) as an inverted plastic flow rule.

Viewing (2.4) as a plastic flow rule suggests the following terminology: The material exhibits thermal softening at state variables (θ,γ,p), where $f_\theta(\theta,\gamma,p) < 0$, strain hardening at state variables where $f_\gamma(\theta,\gamma,p) > 0$, and strain softening when $f_\gamma(\theta,\gamma,p) < 0$. The amounts of the slopes of f in the directions θ, γ, and p measure the degree of thermal softening, strain hardening (or softening), and strain-rate sensitivity, respectively. The difficulty of performing high strain-rate experiments causes uncertainty as to the specific form of the constitutive relation (2.4). Examples that have been extensively used are the power law or the Arrhenius law outlined later.

2.2. Nondimensionalization. To turn the system into a dimensionless form we introduce the following nondimensional variables:

$$\hat{x} = \frac{x}{d}, \qquad \hat{t} = t\dot{\gamma}_0, \qquad \hat{v} = \frac{v}{V}, \qquad \hat{\theta} = \frac{\theta}{\tau_0/\rho c}, \qquad \hat{\sigma} = \frac{\sigma}{\tau_0}, \tag{2.7}$$

where we introduce a nominal stress τ_0 to be appropriately selected later, a nominal temperature $\theta_0 = \frac{\tau_0}{\rho c}$, and a nominal strain rate $\dot{\gamma}_0 = \frac{V}{d}$.

With these choices we obtain the nondimensional system

$$\begin{aligned} \hat{v}_{\hat{t}} &= \frac{1}{r}\ \hat{\sigma}_{\hat{x}}, \\ \hat{\theta}_{\hat{t}} &= \kappa \hat{\theta}_{\hat{x}\hat{x}} + \beta \hat{\sigma}\ \hat{v}_{\hat{x}}, \\ \hat{\gamma}_{\hat{t}} &= \hat{v}_{\hat{x}}, \end{aligned} \tag{2.8}$$

where the nondimensional numbers are

$$r = \frac{\rho V^2}{\tau_0}, \qquad \kappa = \frac{k}{\rho c V d}, \tag{2.9}$$

while β is nondimensional by its very nature. The number r is a ratio of inertial to viscoplastic stresses and depends on the choice of the normalizing stress τ_0. The constitutive law (2.4) turns to the nondimensional form

$$\hat{\sigma} = \hat{f}(\hat{\theta}, \hat{\gamma}, \hat{\gamma}_{\hat{t}}) = \frac{1}{\tau_0} f\left(\frac{\tau_0}{\rho c}\hat{\theta}, \hat{\gamma}, \dot{\gamma}_0 \hat{\gamma}_{\hat{t}}\right).$$

The freedom in the choice of τ_0 is useful in normalizing the form of $\hat{f}$.

2.3. Power laws. In the experimental literature on shear bands at high strain-rates there is extensive use of constitutive laws in the form of power laws (e.g., [15], [14])

$$\sigma = G\ \left(\frac{\theta}{\theta_r}\right)^{-\alpha}\ \left(\frac{\gamma}{\gamma_r}\right)^{m}\ \left(\frac{\gamma_t}{\dot{\gamma}_r}\right)^{n} = G_0\ \theta^{-\alpha}\gamma^m\gamma_t^n. \tag{2.10}$$

Here, α, m, n denote the thermal softening, strain hardening, and strain-rate sensitivity parameters, respectively, G is a material constant, and $\theta_r,\ \gamma_r, \dot{\gamma}_r$ are some reference values for temperature, strain, and strain rate, respectively. Specifically, $\gamma_r \simeq 0.01$ is the strain at yield in a quasi-static simple shear test at a nominal strain rate $\dot{\gamma} = 10^{-4}/s$ for most steels. There is no unique choice for the other reference values, but the simplest choices are $\dot{\gamma}_r = 10^3/s,\ \theta_r = 300K$. This corresponds to the nominal strain rate and ambient temperature of the usual torsional experiment. In (2.10) θ and θ_r are measured in Kelvin. This power law model has been used extensively to model steels that exhibit shear bands [15], [14] and is entirely empirical, but it allows considerable flexibility in fitting experimental data over an extended range. According to experimental data for most steels we have $\alpha = O(10^{-1}),\ m = O(10^{-2})$, and $n = O(10^{-2})$.

The freedom in the choice of the nominal stress τ_0 is useful to simplify the form of $\hat{f}$. For (2.10), if we select τ_0 such that

$$\frac{1}{\tau_0} G\ \left(\frac{\frac{\tau_0}{\rho c}}{\theta_r}\right)^{-\alpha}\ \left(\frac{1}{\gamma_r}\right)^{m}\ \left(\frac{\dot{\gamma}_0}{\dot{\gamma}_r}\right)^{n} = 1,$$

it yields the nondimensional form $\hat{\sigma} = \hat{\theta}^{-\alpha}\hat{\gamma}^m\hat{\gamma}_t^n$.

2.4. The mathematical model. We collect the nondimensional form of the equations, dropping the hats, in the form

$$\begin{aligned} v_t &= \frac{1}{r}\,\sigma_x, \\ \theta_t &= \kappa\theta_{xx} + \sigma\gamma_t, \\ \gamma_t &= v_x, \\ \sigma &= f(\theta,\gamma,\gamma_t), \end{aligned} \tag{2.11}$$

where r, κ are given by (2.9) and we have taken the (not so important) constant $\beta = 1$. For the stress, we use the empirical power law, which, for the appropriate choice of τ_0, takes the normalized form

$$\sigma = \theta^{-\alpha}\gamma^m\gamma_t^n. \tag{2.12}$$

The parameters $\alpha > 0$, m and $n > 0$ serve as measures of the degree of thermal softening, strain hardening (or softening), and strain-rate sensitivity. Another commonly used constitutive relation is the Arrhenius law

$$\sigma = e^{-\alpha\theta}\gamma_t^n. \tag{2.13}$$

In the form (2.13) the Arrhenius law does not exhibit any strain hardening, and the parameters α and n measure the degree of thermal softening and strain-rate sensitivity, respectively. The boundary conditions are prescribed velocities at the ends of the plates, in the nondimensional form

$$v(0,t) = 0, \quad v(1,t) = 1, \quad t \geq 0, \tag{2.14}$$

and thermal insulation at the two ends

$$\theta_x(0,t) = 0, \quad \theta_x(1,t) = 0, \quad t \geq 0. \tag{2.15}$$

We impose initial conditions

$$v(x,0) = v_0(x),\ \theta(x,0) = \theta_0(x) > 0,\ \gamma(x,0) = \gamma_0(x) > 0, \quad x \in [0,1]. \tag{2.16}$$

For the initial data, we take $v_{0x} > 0$, in which case a maximum principle shows that $\gamma_t = v_x > 0$ at all times, and thus all powers are well defined.

2.5. Isothermal versus adiabatic deformations. To illustrate the effect of thermal softening on *spatially uniform* deformations, the isothermal and adiabatic cases are contrasted. Consider a deformation where the plate is subjected to steady shearing, with boundary velocities $v = 0$ at $x = 0$ and $v = 1$ at $x = 1$.

(i) In an *isothermal* deformation the temperature is kept constant, say θ_0, by appropriately removing the produced heat due to the plastic work. The "measured" stress-strain response in this idealized situation coincides with the $\sigma - \gamma$ graph of the function $\sigma = f(\theta_0, \gamma, 1)$. The slope of the graph is measured by $f_\gamma(\theta_0, \gamma, 1)$, and, for a strain-hardening material, the graph $\sigma - \gamma$ is monotonically increasing.

(ii) The situation in an *adiabatic* deformation is understood by studying a special class of solutions describing *uniform shearing.* These are

$$\begin{aligned} v_s(x,t) &= x, \\ \gamma_s(x,t) &= \gamma_s(t) = t + \gamma_0, \\ \theta_s(x,t) &= \theta_s(t), \end{aligned}$$

where

$$\begin{aligned} \frac{d\theta_s}{dt} &= f(\theta_s, t+\gamma_0, 1), \\ \theta_s(0) &= \theta_0, \end{aligned}$$

and γ_0, θ_0 are positive constants, standing for the initial values of the strain and temperature. The resulting stress is given by the graph of the function

$$\sigma_s(t) = f(\theta_s(t), t+\gamma_0, \dot{\gamma}_0),$$

which may be interpreted as stress versus time but also as stress versus (average) strain. This effective stress-strain curve coincides with the $\sigma_s - t$ graph, and the material exhibits effective hardening in the increasing parts of the graph and effective softening in the decreasing parts. The slope is determined by the sign of the quantity $f_\theta\, f + f_\gamma$, when this sign is negative, the combined effect of strain hardening and thermal softening delivers net softening. For instance, for a strain-hardening ($m > 0$) power law (2.12), the uniform shearing solution reads

$$\begin{aligned} \gamma_s(t) &= t+\gamma_0, \\ \theta_s(t) &= \left[\theta_0^{1+\alpha} + \frac{1+\alpha}{m+1}\left[(t+\gamma_0)^{m+1} - \gamma_0^{m+1}\right]\right]^{\frac{1}{1+\alpha}}, \\ \sigma_s(t) &= \theta_s^{-\alpha}(t)(t+\gamma_0)^m, \end{aligned} \tag{2.17}$$

and a simple computation yields

$$\frac{d\sigma_s}{dt} = \theta_s(t)^{-2\alpha-1}\gamma_s(t)^{2m}\left[\frac{-\alpha+m}{m+1} + \frac{m}{(t+\gamma_0)^{m+1}}\left[\theta_0^{1+\alpha} - \frac{1+\alpha}{m+1}\gamma_0^{m+1}\right]\right].$$

For parameter values ranging in the region $m > \alpha$, the graph $\sigma_s(t)$-t is increasing, and the material exhibits net hardening. By contrast, for parameter values ranging in the region $m < \alpha$, $\sigma_s(t)$ may initially increase but eventually decreases with t. In this range, the combined effect of thermal softening and strain-hardening results to net softening, and it is precisely this effect that is considered as a necessary (though not sufficient) cause of the shear band formation process.

2.6. Strain softening versus strain-rate sensitivity. It is generally maintained that strain softening has a destabilizing influence, tending to amplify small nonuniformities. To illustrate the nature of the instability, consider the model

$$\begin{aligned} v_t &= \tau(\gamma)_x, \\ \gamma_t &= v_x, \end{aligned} \tag{2.18}$$

with $\tau'(u) < 0$. This model describes isothermal motions of a strain-softening, inelastic material. The system (2.18) is elliptic in the t-direction, and the initial value problem is ill-posed. The uniform shearing solution

$$\hat{v} = x, \quad \hat{\gamma} = t+\gamma_0,$$

γ_0 constant, is still a special class of solutions to this problem.

By contrast, strain-rate dependence tends to diffuse nonuniformities in the strain-rate and/or the stress, and it may hinder or even altogether suppress instability. That is confirmed, for example, by considering the system

$$\begin{aligned} v_t &= (\tau(\gamma)v_x^n)_x, \\ \gamma_t &= v_x, \end{aligned} \tag{2.19}$$

with $\tau(u) > 0$ and $\tau'(u) < 0$. This system is a parabolic regularization of the elliptic problem (2.18), and it is precisely the competition between an ill-posed equation and a regularizing effect that is hidden behind the shear band formation problem. If one considers the linearization of the uniform shear solution

$$v = x + \hat{V}, \quad \gamma = t + \gamma_0 + \hat{\Gamma}, \tag{2.20}$$

we see that the linearized problem from $(\hat{V}, \hat{\Gamma})$ reads

$$\begin{aligned} \hat{V}_t &= n\tau(t+\gamma_0)\hat{V}_{xx} + \tau'(t+\gamma_0)\hat{\Gamma}_x, \\ \hat{\Gamma}_t &= \hat{V}_x. \end{aligned} \tag{2.21}$$

The form of the linearized problem is a parabolic regularization of an elliptic initial value problem and indicates that strain-rate sensitivity provides a stabilizing effect to the destabilizing mechanism of strain softening.

To quantify the role of the various effects—thermal softening, strain-hardening, strain-rate sensitivity, and heat diffusion—at the level of the linearized problem, Molinari and Clifton [17] suggest the notion that the uniform shearing solution is stable if the perturbation of the uniform shearing solution grows slower than the basic solution (2.17) and is unstable if the perturbation grows faster than the solution (2.17). It has been conjectured in [17], based on linearized analysis of such "relative perturbations" and some additional plausibility arguments, that, for power laws, the uniform shearing solution is stable in the parameter range $q = m + n - \alpha > 0$ and unstable in the complementary region $q = m + n - \alpha < 0$. The relative perturbation analysis is not straightforward to rigorously justify, as it requires stability analysis for nonautonomous systems. Nevertheless, the linearized analysis was carried out for (2.21) using maximum principles (see [22]), and, in this case, the conjecture was verified.

Nonlinear analysis has been more efficient for providing stability results and validating the above criterion in various special cases [9, 18, 19, 21]. A complete understanding for the full model exists only in the case of stress boundary conditions which are energetically fairly demanding: it is shown in [20] that unstable response and formation of shear bands occurs in certain parameter regimes and that the process of shear band formation is concurrent with a collapse of the stress diffusion mechanisms of the material. The case of velocity boundary conditions is energetically more benign and closer to the experimental setup. For this case, the only available instability results concern a temperature-dependent Newtonian fluid [3] (or a strain softening rate-sensitive solid [22]) and indicate that a large perturbation of the temperature (or the strain) can lead to localization and formation of bands at large times. However, a precise quantification of the onset of localization is at present unavailable and will be pursued in later sections of this work.

2.7. Effect of thermal diffusion. Although the early deformation can with no considerable error be regarded as adiabatic, when localization sets in and temperature gradients across a band become very large, thermal diffusion effects can no longer be regarded as negligible. The morphology of a fully formed band in the late stages of deformation is thus influenced by heat conduction balancing the heat production from plastic work. In [26], an extensive numerical treatment of fully developed shear bands, Walter noticed that, due to heat conduction, the strain rate essentially becomes independent of time in late stages of deformation, even though the temperature and stress continue to evolve. We refer to [16, 10, 12] for studies of the effect of heat conduction in a variety of models.

3. Adiabatic shear. We consider now the adiabatic form ($\kappa = 0$) of the nondimensional system (2.11) with a power law stress

$$\begin{aligned} v_t &= \frac{1}{r}\ \sigma_x, \\ \theta_t &= \sigma \gamma_t, \\ \gamma_t &= v_x, \\ \sigma &= \theta^{-\alpha} \gamma^m \gamma_t^n. \end{aligned} \tag{3.1}$$

In this section we outline various formulations of the problem that are useful in what follows.

Stress formulation. There is a reformulation of the problem (3.1), in the form of a reaction-diffusion system, that has been quite instructive in the development of shear band theory (see [20]). Suppose that σ, θ, γ are considered the independent variables. A simple but lengthy computation shows that they satisfy the reaction-diffusion system

$$\begin{aligned} \sigma_t &= \frac{n}{r}\ \theta^{-\frac{\alpha}{n}}\ \gamma^{\frac{m}{n}}\ \sigma^{\frac{n-1}{n}} \sigma_{xx} + \left(-\alpha \frac{\sigma}{\theta} + \frac{m}{\gamma} \right) \theta^{\frac{\alpha}{n}}\ \gamma^{-\frac{m}{n}}\ \sigma^{\frac{n+1}{n}}, \\ \gamma_t &= \theta^{\frac{\alpha}{n}}\ \gamma^{-\frac{m}{n}}\ \sigma^{\frac{1}{n}}, \\ \theta_t &= \theta^{\frac{\alpha}{n}}\ \gamma^{-\frac{m}{n}}\ \sigma^{\frac{n+1}{n}}. \end{aligned} \tag{3.2}$$

Conversely, given a solution (σ, θ, γ) of (3.2), if we define v_x by

$$v_x = \theta^{\frac{\alpha}{n}}\ \gamma^{-\frac{m}{n}}\ \sigma^{\frac{1}{n}},$$

then (v, θ, γ) satisfies (3.1).

Time rescaling. Motivated by the form of the uniform shearing solutions (2.17), one may introduce a rescaling of the dependent variables and time in the following form:

$$\begin{aligned} &\theta(x,t) = (t+1)^{\frac{m+1}{\alpha+1}} \Theta(x, \tau(t)), \quad \gamma(x,t) = (t+1)\Gamma(x, \tau(t)), \\ &\sigma(x,t) = (t+1)^{\frac{m-\alpha}{\alpha+1}} \Sigma(x, \tau(t)), \quad v(x,t) = V(x, \tau(t)), \quad \tau = \ln(1+t). \end{aligned} \tag{3.3}$$

In the new variables $(V, \Theta, \Gamma, \Sigma)$ the system (3.1) becomes

$$\begin{aligned} V_\tau &= \frac{1}{r} e^{\frac{m+1}{1+\alpha}\tau}\ \Sigma_x, \\ \Gamma_\tau &= V_x - \Gamma, \\ \Theta_\tau &= \Sigma V_x - \frac{m+1}{1+\alpha} \Theta, \\ \Sigma &= \Theta^{-\alpha} \Gamma^m V_x^n. \end{aligned} \tag{3.4}$$

Accordingly, the system (3.2) takes the form

$$
\begin{aligned}
\Sigma_\tau &= \frac{n}{r}\, e^{\frac{m+1}{1+\alpha}\tau}\Theta^{-\frac{\alpha}{n}}\ \Gamma^{\frac{m}{n}}\ \Sigma^{\frac{n-1}{n}}\Sigma_{xx} \\
&\qquad\qquad + \left(-\alpha\frac{\Sigma}{\Theta} + \frac{m}{\Gamma}\right)\Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{n+1}{n}} + \Sigma\frac{\alpha - m}{1+\alpha}, \\
\Gamma_\tau &= \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - \Gamma, \\
\Theta_\tau &= \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{n+1}{n}} - \frac{m+1}{1+\alpha}\Theta.
\end{aligned}
\tag{3.5}
$$

Various properties of (3.5) will be noted in forthcoming sections. To understand its usefulness, note that the rescaled variants of the uniform shearing solutions (2.17), given by

$$
\theta_s(t) = (t+1)^{\frac{m+1}{\alpha+1}}\Theta_s(\tau(t)), \quad \gamma_s(t) = (t+1)\Gamma_s(\tau(t)), \quad \sigma_s(t) = (t+1)^{\frac{m-\alpha}{\alpha+1}}\Sigma_s(\tau(t)),
$$

have the long-time behavior

$$
\begin{aligned}
&\Theta_s(\tau) \to \left(\frac{1+\alpha}{1+m}\right)^{\frac{1}{1+\alpha}}, &&\Gamma_s(\tau) \to 1, \\
&\Sigma_s(\tau) \to \left(\frac{1+\alpha}{1+m}\right)^{-\frac{\alpha}{1+\alpha}}, &&\Sigma_s^{\frac{1}{n}}\Theta_s^{\frac{\alpha}{n}}\Gamma_s^{-\frac{m}{n}} \to 1,
\end{aligned}
\tag{3.6}
$$

as $\tau \to \infty$ independently of the values of the initial constants θ_0, γ_0.

4. Non-Newtonian fluids with temperature-dependent viscosity. Various simplified models have been used in the mathematical and mechanics literature of shear band formation. One example is models that neglect thermal softening like (2.19). Another class neglects the effect of strain hardening ($m = 0$) in which case the kinematic compatibility equation decouples, and the system consists of two equations. In the adiabatic case ($\kappa = 0$), the resulting model reads

$$
\begin{aligned}
v_t &= \frac{1}{r}\sigma_x, \\
\theta_t &= \sigma\, v_x,
\end{aligned}
\tag{4.1}
$$

with the power law

$$
\sigma = \theta^{-\alpha} v_x^n, \tag{4.2}
$$

and may be viewed as describing a non-Newtonian fluid with temperature-dependent viscosity. The problem is set in $[0, 1]$ with velocity boundary conditions (2.14), and the objective is to examine the stability of uniform shearing flows (2.17). In this context, the question becomes whether the destabilizing effect of the decreasing and spatially nonhomogeneous viscosity is sufficiently powerful to overcome the stabilizing influence of momentum diffusion and induce localization of shear. We introduce the change of variables

$$
\begin{aligned}
&\theta(x,t) = (t+1)^{\frac{1}{\alpha+1}}\Theta(x,\tau(t)), \quad \sigma(x,t) = (t+1)^{-\frac{\alpha}{\alpha+1}}\Sigma(x,\tau(t)), \\
&v(x,t) = V(x,\tau(t)), \quad \tau = \ln(1+t),
\end{aligned}
\tag{4.3}
$$

(analogous to (3.3)) and obtain the rescaled system

$$\begin{aligned}\Sigma_\tau &= \frac{n}{r}\, e^{\frac{1}{1+\alpha}\tau}\Theta^{-\frac{\alpha}{n}}\ \Sigma^{\frac{n-1}{n}}\Sigma_{xx} + \left(-\alpha\Theta^{\frac{\alpha}{n}-1}\Sigma^{\frac{n+1}{n}} + \frac{\alpha}{1+\alpha}\right)\Sigma,\\ \Theta_\tau &= \left(\Theta^{\frac{\alpha}{n}-1}\ \Sigma^{\frac{n+1}{n}} - \frac{1}{1+\alpha}\right)\Theta.\end{aligned} \tag{4.4}$$

(V, Θ, Σ) also satisfy the equations

$$\Sigma = \Theta^{-\alpha} V_x^n, \tag{4.5}$$

$$V_\tau = \frac{1}{r} e^{\frac{1}{1+\alpha}\tau}\Sigma_x. \tag{4.6}$$

The rescaled variants of the uniform shearing solution (2.17) enjoy the asymptotic behavior

$$\begin{aligned}&\Theta_s(\tau) \to (1+\alpha)^{\frac{1}{1+\alpha}},\\ &\Sigma_s(\tau) \to (1+\alpha)^{-\frac{\alpha}{1+\alpha}},\\ &\Sigma_s^{\frac{1}{n}}\Theta_s^{\frac{\alpha}{n}} \to 1 \qquad\qquad \text{as } \tau \to \infty.\end{aligned}$$

In this section we discuss the stability properties of the uniform shearing solutions in the model (4.4).

4.1. Equilibria and orbits. Consider the reaction part of the system (4.4), i.e., the associated system of ordinary differential equations

$$\begin{aligned}\Sigma_\tau &= -\alpha\Sigma\left(\Theta^{\frac{\alpha}{n}-1}\Sigma^{\frac{n+1}{n}} - \frac{1}{1+\alpha}\right),\\ \Theta_\tau &= \Theta\left(\Theta^{\frac{\alpha}{n}-1}\ \Sigma^{\frac{n+1}{n}} - \frac{1}{1+\alpha}\right).\end{aligned} \tag{4.7}$$

We observe that

$$\frac{\Sigma_\tau}{\Sigma} + \alpha\frac{\Theta_\tau}{\Theta} = 0 \Longleftrightarrow \partial_\tau\left(\frac{\Sigma}{\Theta^{-\alpha}}\right) = 0,$$

which implies that $\Sigma\ \Theta^\alpha$ is constant along the orbits of (4.7). The equilibria of (4.7) are located on the curve

$$\Sigma = \left(\frac{1}{1+\alpha}\right)^{\frac{n}{n+1}}\Theta^{\frac{n-\alpha}{n+1}}. \tag{4.8}$$

The curve of equilibria changes monotonicity depending on the sign of

$$q = -\alpha + n; \tag{4.9}$$

it is increasing for $q > 0$ and decreasing for $q < 0$. In Figure 4.1 the orbits of the system (4.7) along with the curve (4.8) are presented in the case $q > 0$ as well as for the case $q < 0$. The point

$$(\Theta_m, \Sigma_m) = ((1+\alpha)^{\frac{1}{1+\alpha}}, (1+\alpha)^{-\frac{\alpha}{1+\alpha}}), \tag{4.10}$$

see Figure 4.1, corresponds to the asymptotic state of uniform shear in the rescaled variables. When the uniform shearing solution (3.6) is asymptotically stable, then trajectories of the system (4.4) should approach the point (Θ_m, Σ_m) as $\tau \to \infty$.

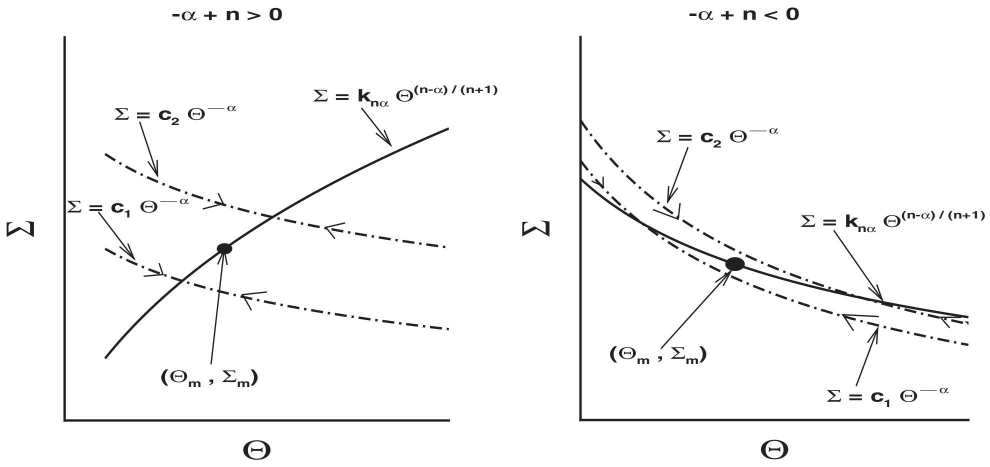


FIG. 4.1. *Orbits of ODE system* (4.7), $k_{n\alpha} = \left(\frac{1}{\alpha+1}\right)^{\frac{n}{n+1}}$.

4.2. Invariant regions, the stable regime. In the case $q = -\alpha + n > 0$, any rectangle

$$[\Theta_-, \Theta_+] \times [\Sigma_-, \Sigma_+], \quad \text{with } (\Theta_-, \Sigma_-),\ (\Theta_+, \Sigma_+) \text{ equilibria},$$

is invariant under the flow of (4.7). The theory of invariant regions for parabolic systems [6] then implies that such rectangles are also invariant under the flow of the reaction-diffusion system (4.4). For initial data taking values in one of the invariant rectangles

$$\begin{aligned} \Theta_- &< \Theta_0(x) < \Theta_+, \\ \Sigma_- &< \Sigma_0(x) < \Sigma_+, \end{aligned}$$

the solutions of (4.4) satisfy

$$\begin{aligned} \Theta_- &< \Theta(x,\tau) < \Theta_+, \\ \Sigma_- &< \Sigma(x,\tau) < \Sigma_+, \\ \Sigma_-^{\frac{1}{n}} \Theta_-^{\frac{\alpha}{n}} &< V_x(x,\tau) = \Sigma^{\frac{1}{n}}(x,\tau) \Theta^{\frac{\alpha}{n}}(x,\tau) < \Sigma_+^{\frac{1}{n}} \Theta_+^{\frac{\alpha}{n}}. \end{aligned} \tag{4.11}$$

The reader should note that the invariant regions property is lost in the complementary region $q < 0$, and no rectangle of the form $[\Theta_-, \Theta_+] \times [\Sigma_-, \Sigma_+]$ is invariant for the reaction system (4.7). The bounds (4.11) yield time-dependent estimates for solutions of (4.1), (4.2):

$$\begin{aligned} \Theta_-(t+1)^{\frac{1}{1+\alpha}} &< \theta(x,t) < \Theta_+ t + 1)^{\frac{1}{1+\alpha}}, \\ \Sigma_-(t+1)^{\frac{-\alpha}{1+\alpha}} &< \sigma(x,t) < \Sigma_+(t+1)^{\frac{-\alpha}{1+\alpha}}, \\ \Sigma_-^{\frac{1}{n}} \Theta_-^{\frac{\alpha}{n}} &< v_x(x,t) < \Sigma_+^{\frac{1}{n}} \Theta_+^{\frac{\alpha}{n}}. \end{aligned} \tag{4.12}$$

These are used to show that the uniform shear solution is asymptotically stable.

THEOREM 4.1. *Let* $q = -\alpha + n > 0$, *and consider a solution* (v, θ) *of* (4.1), (4.2), *with initial data* $\theta_0(x) > 0$ *and* $\sigma_0(x) > 0$. *Then* (v, θ) *is defined for all times and has* *the asymptotic behavior*

$$v_x(x,t) = 1 + O\left((t+1)^{-\frac{n-\alpha}{n(1+\alpha)}}\right), \tag{4.13}$$

$$\theta(x,t) = \Theta_m (t+1)^{\frac{1}{1+\alpha}}\left(1 + O\big((t+1)^{-\frac{n-\alpha}{n(1+\alpha)}}\big)\right), \tag{4.14}$$

$$\sigma(x,t) = \Sigma_m (t+1)^{\frac{-\alpha}{1+\alpha}}\left(1 + O\big((t+1)^{-\frac{n-\alpha}{n(1+\alpha)}}\big)\right), \tag{4.15}$$

as $t \to \infty$, where (Θ_m, Σ_m) are given in (4.10).

Theorem 4.1 was proved in [9, 18] using detailed energy estimates to derive the time-dependent bounds (4.12). These estimations are considerably simplified using the invariant regions presented above. The remainder of the asymptotic stability proof is presented in the appendix.

4.3. The unstable regime. The stability of the uniform shearing solution in the complementary region $q = -\alpha + n < 0$ is at present unknown. In fact, it is even unknown whether solutions exist globally in time or, in contrast, blow up in finite time. Numerical investigations indicate development of shear bands in this regime. In addition, there are two theoretical results that are also backing this direction. First, consider initial data $\theta_0(x)$, $v_0(x)$ such that

$$v_0(x) = x, \tag{4.16}$$

$$\theta_0(x) = \begin{cases} \bar{\theta} & x \notin I_\delta, \\ U(x) & x \in I_\delta, \end{cases} \tag{4.17}$$

where $\bar{\theta}$ is a constant, $I_\delta = (y-\delta, y+\delta)$ is a (small) interval centered around a given point $y \in (0,1)$, and $U(x)$ is the initial temperature profile in I_δ and is selected so that θ_0 is smooth.

Theorem 4.2. *Let $q < 0$ and (v, θ) be a solution of* (4.1), (4.2) *with initial data* (4.16), (4.17). *If $U(y)$ is selected sufficiently large, then either the solution blows up in finite time,*

$$\limsup_{t \to T^*} \sup_{x \in [0,1]} \theta(x,t) \to \infty \quad \textit{for some } T^* < \infty,$$

or else $T^* = \infty$ and (v, θ) has the asymptotic behavior

$$v(x,t) = \begin{cases} 0 + O\left((t+1)^{-\frac{1}{n+1}}\right) & x < y - \delta, \\ 1 + O\left((t+1)^{-\frac{1}{n+1}}\right) & y + \delta < x, \end{cases} \tag{4.18}$$

and $\theta(x,t)$ approaches a limiting temperature profile for $x \notin I_\delta$ as $t \to \infty$.

Theorem 4.2 was proved in [3] for the case of a Newtonian fluid ($n = 1$). We extend this result for non-Newtonian temperature-dependent flows. The proof is provided in the appendix, and it yields a quantitative criterion for a size of an initial temperature perturbation $U(x)$ that suffices to induce instability. The reader can check that even a moderate perturbation will suffice, but an arbitrarily small perturbation of the uniform temperature is excluded. The question remains, what is the basic mechanism that induces this instability? This question is answered at the level of the full system (3.1) in the following two sections, of course, including the special case (4.1), (4.2).

5. The long-time response of adiabatic shear as a relaxation limit. Our next goal is to derive an effective equation describing the long-time behavior of the system (3.1) by using convenient scaling limits. This is done in three steps: First, we consider the system (3.5) or the equivalent form (3.4) and point out certain analogies with the structure of relaxation systems (like equilibrium manifolds, moment equations). Then we consider a modified version of the system (3.4) and show how to introduce a scaled limit that describes the long-time response and how to compute the effective response by a process analogous to the Chapman–Enskog expansion. The analysis applies to the modified system (see (5.9)) which shares the same general structure as (3.5) but also has an important difference. Then in section 6, we consider the original system (3.5) and modify the change of variables (3.3), keeping in mind that we calculate perturbed profiles of time-dependent solutions. An analogous procedure then leads to the effective equation describing the long-time response of (3.1).

5.1. Some analogies to the theory of relaxation systems. We first point out certain analogies between (3.4) and the theory of relaxation processes. Consider a solution (Σ, Θ, Γ) of (3.5), and note that it satisfies the identity

$$\frac{\Sigma_\tau}{\Sigma} - m\frac{\Gamma_\tau}{\Gamma} + \alpha\frac{\Theta_\tau}{\Theta} = \frac{n}{r}\, e^{\frac{m+1}{1+\alpha}\tau}\Theta^{-\frac{\alpha}{n}}\ \Gamma^{\frac{m}{n}}\ \Sigma^{-\frac{1}{n}}\Sigma_{xx}, \tag{5.1}$$

that is, $U = (\Sigma\ \Theta^{\alpha}\ \Gamma^{-m})^{\frac{1}{n}}$ satisfies a conservation law

$$\partial_\tau U = \frac{1}{r}\, e^{\frac{m+1}{1+\alpha}\tau}\ \Sigma_{xx}. \tag{5.2}$$

Equation (5.2) is precisely the first equation in (3.4), and it may be interpreted as a conservation law for the quantity $U = V_x$ that arises as a moment equation for the reaction-diffusion system (3.5). The reaction system associated to (3.5) is

$$\begin{aligned}
\Sigma_\tau &= -\alpha\Sigma\left(\frac{\Sigma}{\Theta}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - \frac{m+1}{1+\alpha}\right) + m\Sigma\left(\frac{1}{\Gamma}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - 1\right),\\
\Gamma_\tau &= \Gamma\left(\frac{1}{\Gamma}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - 1\right),\\
\Theta_\tau &= \Theta\left(\frac{\Sigma}{\Theta}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - \frac{m+1}{1+\alpha}\right).
\end{aligned} \tag{5.3}$$

For the initial data, we assume that $\Gamma_0(x) > 0$, $\Theta_0(x) > 0$, $\Sigma_0(x) > 0$, and they depend parametrically on x.

Equilibria. The equilibria of (5.3) are the solutions of the algebraic system

$$\begin{aligned}
&\frac{1}{\Gamma}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - 1 = 0,\\
&\frac{\Sigma}{\Theta}\ \Theta^{\frac{\alpha}{n}}\ \Gamma^{-\frac{m}{n}}\ \Sigma^{\frac{1}{n}} - \frac{m+1}{1+\alpha} = 0
\end{aligned}$$

or equivalently,

$$\frac{\Sigma\Gamma}{\Theta} = \frac{m+1}{1+\alpha}, \quad \Sigma = \Theta^{-\alpha}\ \Gamma^{m+n}. \tag{5.4}$$

The equilibria form a one-parameter family, determined in terms of a parameter $U \in \mathbb{R}^+$, by the equations

$$\begin{aligned}
&\Gamma = U, \\
&\Theta = \left(\frac{1+\alpha}{m+1} \right)^{\frac{1}{1+\alpha}} U^{\frac{m+n+1}{1+\alpha}}, \\
&\Sigma = \left(\frac{1+\alpha}{m+1} \right)^{-\frac{\alpha}{1+\alpha}} U^{\frac{-\alpha+m+n}{1+\alpha}}.
\end{aligned} \tag{5.5}$$

Orbits. Although the system (5.3) is complex in appearance, its orbits are easily computed due to the property that solutions of (5.3) satisfy the conservation law

$$\frac{\Sigma_\tau}{\Sigma} - m\frac{\Gamma_\tau}{\Gamma} + \alpha\frac{\Theta_\tau}{\Theta} = 0 \iff \partial_\tau \left(\Sigma \ \Theta^\alpha \ \Gamma^{-m} \right) = 0.$$

The quantity $\Sigma \ \Theta^\alpha \ \Gamma^{-m}$ thus remains constant along an orbit,

$$\Sigma \ \Theta^\alpha \ \Gamma^{-m} = U^n = \text{constant in time}, \tag{5.6}$$

with the value of $U = V_x$ determined by the values of the initial data. As a result of (5.6), Γ satisfies the differential equation

$$\frac{\Gamma_\tau}{\Gamma} = \frac{U}{\Gamma} - 1.$$

Since U is constant in time and the initial data $\Gamma_0(x) > 0$ are positive, $\Gamma(x, \tau)$ remains bounded above and below by positive bounds depending on $U(x)$ and $\Gamma_0(x)$ and that

$$\Gamma \to U \qquad \text{as } \tau \to \infty.$$

Furthermore, concerning the dynamics of the system of differential equations (5.3), we have

$$\begin{aligned}
\partial_\tau \left(\frac{\Sigma \ \Gamma}{\Theta} \right) &= \frac{\Sigma \ \Gamma}{\Theta} \left(\frac{\Sigma_\tau}{\Sigma} + \frac{\Gamma_\tau}{\Gamma} - \frac{\Theta_\tau}{\Theta} \right) = \frac{\Sigma \ \Gamma}{\Theta} \Theta^{\frac{\alpha}{n}} \ \Gamma^{-\frac{m}{n}} \ \Sigma^{\frac{1}{n}} \left(\frac{m+1}{\Gamma} - (\alpha+1)\frac{\Sigma}{\Theta} \right) \\
&= -(\alpha+1) \left(\frac{\Sigma \ \Gamma}{\Theta} \right) \frac{U}{\Gamma} \left(\frac{\Sigma\Gamma}{\Theta} - \frac{m+1}{1+\alpha} \right).
\end{aligned}$$

This implies $\left| \frac{\Sigma \ \Gamma}{\Theta} - \frac{m+1}{1+\alpha} \right|$ is a decreasing function of τ and

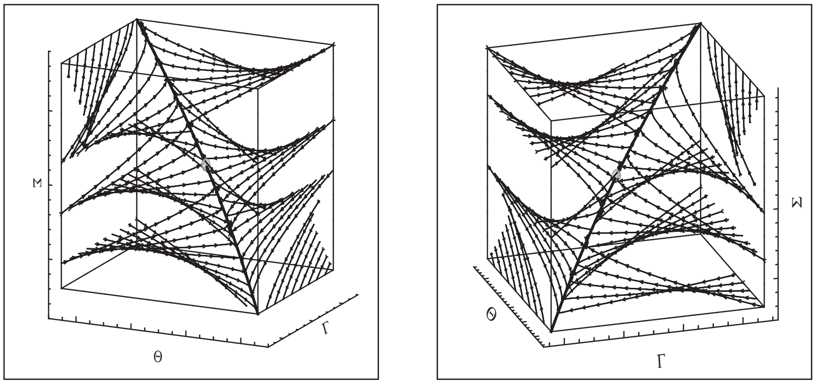
$$\frac{\Sigma \ \Gamma}{\Theta} - \frac{m+1}{1+\alpha} \to 0 \ \text{ as } \tau \to \infty. \tag{5.7}$$

Finally, the identity

$$\begin{aligned}
\frac{\partial_\tau \left(\Sigma^{\frac{1}{n}} \ \Theta^{\frac{\alpha}{n}} \ \Gamma^{-\frac{m+n}{n}} \right)}{\left(\Sigma^{\frac{1}{n}} \ \Theta^{\frac{\alpha}{n}} \ \Gamma^{-\frac{m+n}{n}} \right)} &= \frac{1}{n} \ \frac{\Sigma_\tau}{\Sigma} + \frac{\alpha}{n} \ \frac{\Theta_\tau}{\Theta} - \frac{m+n}{n} \ \frac{\Gamma_\tau}{\Gamma} \\
&= -\left(\Sigma^{\frac{1}{n}} \ \Theta^{\frac{\alpha}{n}} \ \Gamma^{-\frac{m+n}{n}} - 1 \right)
\end{aligned}$$

implies that $\Phi(\tau) = \Sigma^{\frac{1}{n}} \ \Theta^{\frac{\alpha}{n}} \ \Gamma^{-\frac{m+n}{n}}$ satisfies the ordinary differential equation

$$\partial_\tau \Phi = -\Phi \left(\Phi - 1 \right),$$



(a) (b)

FIG. 5.1. *Two views of the flow of the ODE system* (5.3) *in the stable case:* $-\alpha+m+n>0$.

and thus $\Phi(\tau)\to 1$ as $\tau\to\infty$, that is,

$$\Sigma\ \Theta^{\alpha}\ \Gamma^{-(m+n)}\to 1 \quad \text{as } \tau\to\infty. \tag{5.8}$$

We conclude that the orbits of the differential system (5.3) approach the line of equilibria (5.5). Each orbit lies entirely on the surface (5.6), and the specific value of the parameter U is selected by the initial data. Unlike the system with two equations, strong numerical evidence (Figure 5.1) indicates that the system of three equations does not have invariant regions. In particular, Figure 5.1 shows the flow of the vector field generated by (5.3). Notice in the upper-left corner and/or in the lower-right corner of Figure 5.1(a), there are orbits always exiting the box. One can notice the same type of behavior in the bottom and/or the top part of Figure 5.1(b); there are orbits that are no longer confined in a box around the equilibrium manifold.

5.2. Effective equation for a simplified system. Our ultimate goal is to give an effective equation describing the long-time response of the system (3.1). We will present an argument that leads to such an effective equation in the following section. In preparation, we consider a simplified problem that accounts for the main structure of (3.4), by studying a variant where the time dependence of the diffusion term is frozen,

$$\begin{aligned} V_\tau &= \frac{1}{r}\Sigma_x, \\ \Gamma_\tau &= V_x-\Gamma, \\ \Theta_\tau &= \Sigma V_x-\frac{m+1}{1+\alpha}\Theta, \\ \Sigma &= \Theta^{-\alpha}\ \Gamma^m V_x^n. \end{aligned} \tag{5.9}$$

For this modified system, we calculate an effective equation in an asymptotic limit motivated by the theory of relaxation approximations. To this end, set $U=V_x$ in (5.9), and consider a rescaling of time in the form

$$\begin{aligned} &U(x,\tau)=\bar{U}^T\left(x,\frac{\tau}{T}\right), \quad \Theta(x,\tau)=\bar{\Theta}^T\left(x,\frac{\tau}{T}\right), \\ &\Gamma(x,\tau)=\bar{\Gamma}^T\left(x,\frac{\tau}{T}\right), \quad \Sigma(x,\tau)=\bar{\Sigma}^T\left(x,\frac{\tau}{T}\right), \quad s=\frac{\tau}{T}. \end{aligned} \tag{5.10}$$

Then, $(\bar{U}^T, \bar{\Theta}^T, \bar{\Gamma}^T, \bar{\Sigma}^T)$ satisfies the system of equations

$$\begin{aligned} U_s &= \frac{T}{r}\Sigma_{xx}, \\ \Gamma_s &= T(U-\Gamma), \\ \Theta_s &= T\left(\Sigma U - \frac{m+1}{1+\alpha}\Theta\right), \\ \Sigma \ &= \Theta^{-\alpha}\ \Gamma^m U^n, \end{aligned} \tag{5.11}$$

where in (5.11) we dropped the bars to simplify notations. Given a family of solutions of (5.11), we may use the relations

$$\lim_{T\to\infty} U(x,Ts) = \lim_{T\to\infty} \bar{U}^T(x,s)$$

in order to calculate the long-time behavior of solutions of (5.9). Therefore, it suffices to calculate the effective response in the limit $T \to \infty$ of solutions to (5.11). This goal can be achieved by using the procedure of the Chapman–Enskog expansion (e.g., [5]), familiar from the kinetic theory of gases. In order for the conservation law $(5.11)_1$ to provide a nontrivial effective response, we will consider the limit $T \to \infty$, $r \to \infty$ such that $\frac{T}{r} < \infty$ (and for simplicity will take $\frac{T}{r} = 1$). Consider the ansatz for solutions of (5.11):

$$\begin{aligned} \bar{U}^T &= U_0 + \frac{1}{T}U_1 + O\left(\frac{1}{T^2}\right), \quad \bar{\Gamma}^T = \Gamma_0 + \frac{1}{T}\Gamma_1 + O\left(\frac{1}{T^2}\right), \\ \bar{\Theta}^T &= \Theta_0 + \frac{1}{T}\Theta_1 + O\left(\frac{1}{T^2}\right), \quad \bar{\Sigma}^T = \Sigma_0 + \frac{1}{T}\Sigma_1 + O\left(\frac{1}{T^2}\right). \end{aligned} \tag{5.12}$$

Upon introducing the expansions into (5.11) and expanding in orders of T, keeping in mind that $r = T \to \infty$, we obtain consecutively

$$\partial_s\left(U_0 + \frac{1}{T}U_1 + \cdots\right) = \partial_{xx}\left(\Sigma_0 + \frac{1}{T}\Sigma_1 + \cdots\right), \tag{5.13}$$

$$\partial_s\left(\Gamma_0 + \frac{1}{T}\Gamma_1 + \cdots\right) = T\,(U_0 - \Gamma_0) + (U_1 - \Gamma_1) + \cdots, \tag{5.14}$$

$$\partial_s\left(\Theta_0 + \frac{1}{T}\Theta_1 + \cdots\right) = T\left(\Sigma_0 U_0 - \frac{m+1}{1+\alpha}\Theta_0\right) \tag{5.15}$$

$$+\left(\Sigma_1 U_0 + \Sigma_0 U_1 - \frac{m+1}{1+\alpha}\Theta_1\right) + \cdots, \tag{5.16}$$

and finally

$$\begin{aligned} &\Sigma_0 + \frac{1}{T}\Sigma_1 + \cdots \\ &\quad = \left(\Theta_0 + \frac{1}{T}\Theta_1 + \cdots\right)^{-\alpha}\left(\Gamma_0 + \frac{1}{T}\Gamma_1 + \cdots\right)^{m}\left(U_0 + \frac{1}{T}U_1 + \cdots\right)^{n} \\ &\quad = \Theta_0^{-\alpha}\Gamma_0^m U_0^n\left(1 + \frac{1}{T}\left(-\alpha\frac{\Theta_1}{\Theta_0} + m\frac{\Gamma_1}{\Gamma_0} + n\frac{U_1}{U_0}\right) + \cdots\right). \end{aligned} \tag{5.17}$$

Collecting terms of the same order together, we obtain for the 0th order perturbations the equations

$$\begin{aligned} U_0 &= \Gamma_0, \\ \Sigma_0 U_0 &= \frac{m+1}{1+\alpha}\Theta_0, \\ \Sigma_0 &= \Theta_0^{-\alpha}\Gamma_0^m U_0^n, \end{aligned} \tag{5.18}$$

while for the 1st order perturbations, we deduce the equations

$$\begin{aligned} U_1 - \Gamma_1 &= \partial_s \Gamma_0, \\ \Sigma_1 U_0 + \Sigma_0 U_1 - \frac{m+1}{1+\alpha}\Theta_1 &= \partial_s \Theta_0, \\ \left(-\alpha\frac{\Theta_1}{\Theta_0} + m\frac{\Gamma_1}{\Gamma_0} + n\frac{U_1}{U_0}\right)\Sigma_0 &= \Sigma_1, \\ \partial_s U_1 &= \partial_{xx}\Sigma_1. \end{aligned} \tag{5.19}$$

Solving (5.18) we obtain that all 0th order terms can be expressed in terms of the conserved quantity U_0:

$$\begin{aligned} \Gamma_0 &= U_0, \\ \Sigma_0 &= \left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}, \\ \Theta_0 &= \left(\frac{m+1}{1+\alpha}\right)^{-\frac{1}{1+\alpha}} U_0^{\frac{1+m+n}{1+\alpha}}, \\ \partial_s U_0 &= \partial_{xx}\Sigma_0. \end{aligned} \tag{5.20}$$

Note that U_0 is the conserved quantity and that such structure is typical in the theory of relaxation. In addition, we may obtain an evolution equation governing the behavior of the 0th order approximation in closed form as

$$\partial_s U_0 = \partial_{xx}\left(\left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}\right). \tag{5.21}$$

The nature of (5.21) changes, depending on the sign of

$$q = -\alpha + m + n$$

from forward parabolic when $q > 0$ to backward parabolic when $q < 0$. For the parameters ranging in the region $q < 0$, (5.21) is ill-posed, and one needs to derive the next order of the asymptotics. This is accomplished by solving (5.19) for the 1st order approximants $(\Gamma_1, \Theta_1, \Sigma_1)$ and using the expressions (5.20). We then obtain

$$\begin{aligned} \Gamma_1 &= U_1 - \partial_s U_0, \\ \frac{\Sigma_1}{\Sigma_0} &= \frac{-\alpha+m+n}{1+\alpha}\frac{U_1}{U_0} - \frac{m}{1+\alpha}\frac{\partial_s U_0}{U_0} + \frac{\alpha}{1+\alpha}\frac{\partial_s\Theta_0}{\Sigma_0 U_0}, \\ \frac{\Theta_1}{\Theta_0} &= \frac{1+m+n}{1+\alpha}\frac{U_1}{U_0} - \frac{m}{1+\alpha}\frac{\partial_s U_0}{U_0} - \frac{1}{1+\alpha}\frac{\partial_s\Theta_0}{\Sigma_0 U_0}. \end{aligned} \tag{5.22}$$

The corrected form—up to order $O(\frac{1}{T^2})$—of the effective equation is now easily calculated, using (5.20) and (5.22), leads to the equation

$$
\begin{aligned}
\partial_s \left(U_0 + \frac{1}{T} U_1 + O\left(\frac{1}{T^2}\right)\right) &= \partial_{xx}\left(\Sigma_0 + \frac{1}{T}\Sigma_1 + O\left(\frac{1}{T^2}\right)\right) \\
&= \partial_{xx}\Bigg[\left(\left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}\right) \\
&\quad + \frac{1}{T}\Sigma_0\left(\frac{-\alpha+m+n}{1+\alpha}\frac{U_1}{U_0} - \frac{m}{1+\alpha}\frac{\partial_s U_0}{U_0} + \frac{\alpha}{1+\alpha}\frac{\partial_s \Theta_0}{\Sigma_0 U_0}\right) + O\left(\frac{1}{T^2}\right)\Bigg] \\
&= \partial_{xx}\Bigg[\left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}\left(1 + \frac{-\alpha+m+n}{1+\alpha}\frac{U_1}{U_0}\frac{1}{T} + O\left(\frac{1}{T^2}\right)\right) \\
&\quad + \frac{1}{T}\Sigma_0\left(-\frac{m}{1+\alpha}\frac{\partial_s U_0}{U_0} + \frac{\alpha}{1+\alpha}\frac{\partial_s \Theta_0}{\Sigma_0 U_0}\right) + O\left(\frac{1}{T^2}\right)\Bigg].
\end{aligned}
\tag{5.23}
$$

The last objective is to obtain an equation for U that, up to 2nd order, will agree with (5.23). To this end, observe that the first term in the right side of (5.23) satisfies

$$
I_1 = \left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}} U^{\frac{-\alpha+m+n}{1+\alpha}} + O\left(\frac{1}{T^2}\right),
$$

while the second term may be reexpressed using (5.20) and (5.21) as

$$
\begin{aligned}
I_2 = \Sigma_0\left(-\frac{m}{1+\alpha}\frac{\partial_s U_0}{U_0} + \frac{\alpha}{1+\alpha}\frac{\partial_s \Theta_0}{\Sigma_0 U_0}\right) &= \Sigma_0 \frac{\alpha(1+m+n) - m(m+1)}{(m+1)(1+\alpha)}\frac{\partial_s U_0}{U_0} \\
= \frac{\alpha(1+m+n) - m(m+1)}{(m+1)(1+\alpha)}\left(\frac{m+1}{1+\alpha}\right)^{\frac{2\alpha}{1+\alpha}} & U_0^{\frac{-\alpha+m+n}{1+\alpha}-1}\partial_{xx}(U_0^{\frac{-\alpha+m+n}{1+\alpha}}).
\end{aligned}
$$

The effective equation is thus, up to order $O(\frac{1}{T^2})$,

$$
\partial_s U = \partial_{xx}\left(c U^p + \frac{\lambda c^2}{T} U^{p-1}\partial_{xx} U^p\right),
\tag{5.24}
$$

where

$$
p = \frac{-\alpha+m+n}{1+\alpha}, \quad c = \left(\frac{m+1}{1+\alpha}\right)^{\frac{\alpha}{1+\alpha}}, \quad \lambda = \frac{\alpha(1+m+n) - m(m+1)}{(m+1)(1+\alpha)}.
$$

We thus see that the 2nd order approximation acquires an additional effect comprising of a nonlinear fourth order term. When $q = -\alpha+m+n > 0$, this term is a perturbation of a forward parabolic equation. As the forward parabolic term has a stabilizing response, the fourth order term is a small perturbation, and the sign of λ has no effect on the stability properties. By contrast, in the region $q < 0$, the first term provides backward parabolic response, and any stabilization is due only to the fourth order term. Note that

$$
\lambda = \frac{(\alpha - m - n) + n(1+\alpha) + (\alpha - m)m}{(m+1)(1+\alpha)},
$$

and thus when $q = -\alpha + m + n < 0$, it is $\lambda > 0$. The uniform shear solution corresponds to $U = 1$, and, for this reason, we write $U = 1 + u$ and compute the linearized equation for the perturbation u. This reads

$$\partial_s u = c\,p\,\partial_{xx} u + \frac{\lambda c^2}{T}\,p\,\partial_{xxxx} u. \tag{5.25}$$

The Fourier transform of (5.25) satisfies

$$\partial_s \hat{u} = \left(-c\,p\,\xi^2 + \frac{\lambda c^2}{T}\,p\,\xi^4 \right) \hat{u},$$

and thus when $q < 0$, the low frequencies will grow, but the high frequency modes still decay. Hence, for $\lambda > 0$, the linearized equation (5.25) is well posed.

6. Effective response at the onset of localization. We now consider the system (3.1) and will calculate an effective equation for its time response. We consider a modified time-rescaling of the form

$$\begin{aligned} \theta(x,t) &= (t+1)^{\frac{m+1}{\alpha+1}}\,\Theta\left(x, \frac{s(t)}{T}\right), \quad \gamma(x,t) = (t+1)\Gamma\left(x, \frac{s(t)}{T}\right), \\ \sigma(x,t) &= (t+1)^{\frac{m-\alpha}{\alpha+1}}\,\Sigma\left(x, \frac{s(t)}{T}\right), \quad v_x(x,t) = V_x\left(x, \frac{s(t)}{T}\right), \end{aligned} \tag{6.1}$$

where T is a parameter representing a change of time-unit and $s(t) : [0,\infty) \to [0,\infty)$ is to be selected as a monotone increasing, surjective map that represents a change of time-scale. This should be compared to (3.3) that has been used before. Introducing this transformation to (3.1), we obtain the equations

$$\begin{aligned} \partial_s V_x &= \frac{T}{r}\frac{1}{\dot{s}}(t+1)^{\frac{m-\alpha}{1+\alpha}}\,\Sigma_{xx}, \\ (t+1)\frac{\dot{s}}{T}\Theta_s &= \Sigma V_x - \frac{m+1}{1+\alpha}\Theta, \\ (t+1)\frac{\dot{s}}{T}\Gamma_s &= V_x - \Gamma, \\ \Sigma &= \Theta^{-\alpha}\Gamma^m V_x^n. \end{aligned}$$

We select $s(t)$ so that

$$\dot{s} = (t+1)^{\frac{m-\alpha}{1+\alpha}}, \tag{6.2}$$

that is,

$$s(t) = \frac{1}{\beta}\left[(t+1)^{\beta} - 1\right] \iff t(s) = \left(1 + \beta s\right)^{\frac{1}{\beta}} - 1,$$

where $\beta = \frac{m+1}{1+\alpha}$. We are interested in the limit $T \to \infty$, $r \to \infty$ so that $\frac{T}{r} = O(1)$, and, for simplicity, we will take $T = r$. The choice $r = O(T)$ is done for the following reasons: (i) it is expected that the inertial terms play an important role in the shear band formation process, (ii) we wish to retain both terms of the momentum equation at $O(1)$ as $T \to \infty$ (see also Remark 6.1). With these identifications, we deduce that the scaled functions Θ^T, Σ^T, $U^T = V_x^T$, and Γ^T satisfy

$$\begin{aligned} \partial_s U &= \Sigma_{xx}, \\ \frac{1}{T}(\beta s + 1)\Theta_s &= \Sigma U - \frac{m+1}{1+\alpha}\Theta, \\ \frac{1}{T}(\beta s + 1)\Gamma_s &= U - \Gamma, \\ \Sigma &= \Theta^{-\alpha}\Gamma^m U^n. \end{aligned} \tag{6.3}$$

If in the limit $T \to \infty$ the functions $(U^T, \Gamma^T, \Theta^T, \Sigma^T) \to (U_0, \Gamma_0, \Theta_0, \Sigma_0)$, then the limiting $(U_0, \Gamma_0, \Theta_0, \Sigma_0)$ lies in the "quilibrium" manifold

$$\begin{aligned} \Sigma_0 U_0 &= \beta\Theta_0, \\ U_0 &= \Gamma_0, \\ \Sigma_0 &= \Theta_0^{-\alpha}\Gamma_0^m U_0^n \end{aligned} \tag{6.4}$$

and satisfies the conservation law

$$\partial_s U_0 = \partial_{xx}\Sigma_0. \tag{6.5}$$

The latter can be expressed into closed form and leads to an effective equation for the long-time response, in the form

$$\begin{aligned} \Sigma_0 &= \beta^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}, \\ \partial_s U_0 &= \partial_{xx}\left(\beta^{\frac{\alpha}{1+\alpha}} U_0^{\frac{-\alpha+m+n}{1+\alpha}}\right). \end{aligned} \tag{6.6}$$

Observe that U_0 describes the limiting dynamics of v_x as can be seen by taking the change of variable formula

$$v_x(x, t(\tau T)) = U^T(x, \tau) \tag{6.7}$$

to the limit as the unit-scale $T \to \infty$. Equation (6.6) changes type from forward parabolic for $q > 0$ to backward parabolic for $q < 0$, where $q = -\alpha + m + n$.

Remark 6.1. In order to preserve the moment equation at $O(1)$ in the limit $T \to \infty$, we have used the parameter r, by assuming $r = O(T)$. The same effect can be achieved by using a different scaling: One could alternatively scale the x-variable in the parabolic scaling $x \to \frac{x}{\sqrt{T}}$ and retain $r = O(1)$. This scaling would again preserve the equation $\partial_s U = \partial_{xx}\Sigma$ and lead to the same system (6.3) with a coefficient $r = O(1)$. In return, (6.7) would be replaced by

$$v_x\left(y\sqrt{T}, t(\tau T)\right) = U^T(y, \tau),$$

which indicates that the limit $T \to \infty$ would proceed along parabolic rays.

Finally, we perform the Chapman–Enskog procedure to calculate the next term of the correction. This is entirely analogous to the procedure outlined in section 5.2, and the details are omitted. To this end, the ansatz (5.12) is introduced into (6.3); collecting terms of the same order together, we obtain (6.4), (6.5) at the 0th order (leading to (6.6)) and obtain at the 1st order the equations

$$\begin{aligned} (\beta s + 1)\partial_s\Theta_0 &= \Sigma_0 U_1 + \Sigma_1 U_0 - \beta\Theta_1, \\ (\beta s + 1)\partial_s\Gamma_0 &= U_1 - \Gamma_1, \\ \Sigma_1 &= \Sigma_0\left(-\alpha\frac{\Theta_1}{\Theta_0} + m\frac{\Gamma_1}{\Gamma_0} + n\frac{U_1}{U_0}\right), \end{aligned}$$

together with

$$\partial_s U_1 = \partial_{xx}\Sigma_1. \tag{6.8}$$

Solving the former equations yields

$$\begin{aligned}
\frac{\Gamma_1}{U_0} &= \frac{U_1}{U_0} - (\beta s + 1)\frac{\partial_s U_0}{U_0},\\
\frac{\Sigma_1}{\Sigma_0} &= \frac{-\alpha + m + n}{1+\alpha}\frac{U_1}{U_0} - \frac{m}{1+\alpha}(\beta s + 1)\frac{\partial_s U_0}{U_0} + \frac{\alpha}{1+\alpha}(\beta s + 1)\frac{\partial_s \Theta_0}{\Sigma_0 U_0},\\
\frac{\Theta_1}{\Theta_0} &= \frac{1 + m + n}{1+\alpha}\frac{U_1}{U_0} - \frac{m}{1+\alpha}(\beta s + 1)\frac{\partial_s U_0}{U_0} - \frac{1}{1+\alpha}(\beta s + 1)\frac{\partial_s \Theta_0}{\Sigma_0 U_0}.
\end{aligned}$$

We next introduce the values of U_0, U_1, Σ_0, Σ_1 into the equation

$$\partial_s \left(U_0 + \frac{1}{T}U_1 + O\left(\frac{1}{T^2}\right)\right) = \partial_{xx}\left(\Sigma_0 + \frac{1}{T}\Sigma_1 + O\left(\frac{1}{T^2}\right)\right) \tag{6.9}$$

and regroup the terms following section 5.2 to deduce that U^T satisfies up to order $O\left(\frac{1}{T^2}\right)$ the equation

$$\partial_s U = \partial_{xx}\left(c\,U^p + \frac{\lambda c^2}{T}(\beta s + 1)U^{p-1}\partial_{xx}U^p\right), \tag{6.10}$$

where

$$p = \frac{-\alpha + m + n}{1+\alpha}, \quad \beta = \frac{m+1}{1+\alpha}, \quad c = \beta^{\frac{\alpha}{1+\alpha}}, \quad \lambda = \frac{\alpha(1+m+n) - m(m+1)}{(m+1)(1+\alpha)}.$$

When $q < 0$, the second order term is backward parabolic and has a destabilizing role, at the same time $\lambda > 0$ and the fourth order term offers a stabilizing influence.

System versus effective equation: Numerical comparison. Next, we compare numerically the solution of system (6.3) with (6.10). The effective equation (6.10) is a highly nonlinear fourth order equation whose behavior depends drastically on the sign of p. In the stable case $p > 0$, (6.10) is a forward parabolic equation and a fourth order correction term whose sign depends on the parameter λ. In the stable case, the fourth order term does not have a definite sign, since the parameter λ can be either positive or negative but nevertheless without any essential effect on the qualitative behavior of the effective equation. We compare numerically the solution of the system (6.3) and (6.10) for $T = 1000$ at various time instances. We use a standard finite element method for the spatial discretization, coupled with Newton's method for linearization, while the Crank–Nicolson method is used for time stepping. In this case we expect the solution $U = V_x$ to converge as $t \to \infty$ to the uniform shearing solution (2.17). In Figure 6.1 we see an excellent agreement of the two solutions, especially as t grows.

In the unstable case $p < 0$, the behavior of the effective equation (6.10) changes drastically. The leading second order term has a negative sign, thus the effective equation changes to an unstable backward parabolic type. On the other hand, the parameter λ in the unstable case is positive, thus the fourth order term provides a stabilizing effect. The balance of these two mechanisms is a delicate issue theoretically as well as numerically. In general, the numerical solution of linear or nonlinear fourth

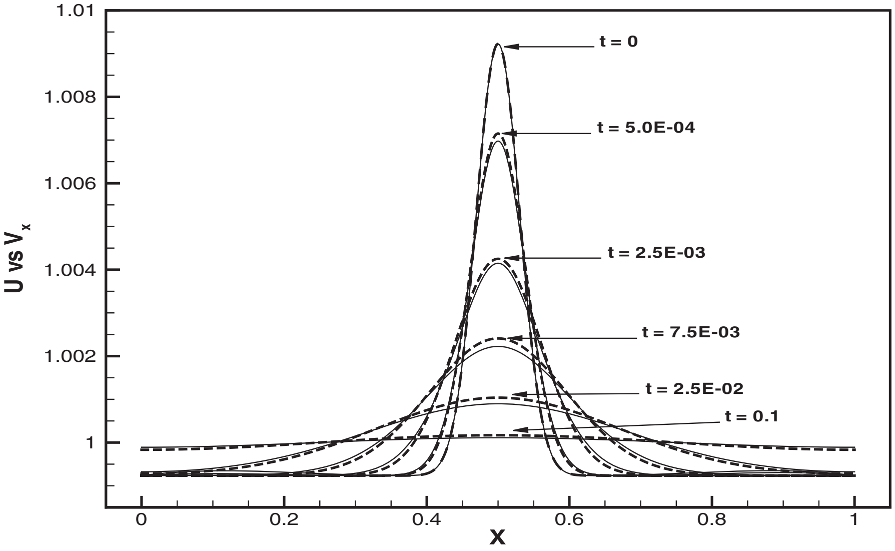


FIG. 6.1. *Comparison in the parameter range* $p > 0$ *of system* (6.3) *(solid line) versus effective equation* (6.9)*(dashed line) for* $T = 1000$.

order equations is not a trivial task. Special numerical techniques have to be applied to capture correctly the behavior of the underlying phenomena. These issues go beyond the scope of the present article and will be the subject of a future work. Detailed numerical results on the behavior of systems (2.11) and the companion system with Fourier heat conduction can be found in [2].

Appendix A. We present in this appendix the proofs of Theorems 4.1 and 4.2. Let (v, θ, σ) be a smooth solution of (4.1), (4.2), and (V, Θ, Σ) the rescaled functions defined in (4.3) and satisfying (4.4), (4.5), and (4.6).

Proof of Theorem 4.1. In the range $q = -\alpha + n > 0$, the system (4.4) is endowed with invariant regions, and (Θ, Σ) and V_x satisfy the bounds (4.11). In what follows, C stands for a generic constant that is independent of time.

Set $g(\tau) = \frac{1}{r} e^{\frac{1}{1+\alpha}\tau}$. Using (4.6), we obtain the identity $\partial_\tau \left(\frac{1}{2g^2} V_\tau^2 \right) = \frac{1}{g} \Sigma_{x\tau} V_\tau$. Integrating by parts over $[0, 1]$ and using $(4.4)_1$, (4.6), and (2.14), we obtain

$$\frac{d}{d\tau} \frac{1}{2} \int_0^1 \frac{1}{g^2} V_\tau^2 dx + \int_0^1 n \Theta^{-\frac{\alpha}{n}} \Sigma^{\frac{n-1}{n}} \frac{1}{g} V_{x\tau}^2 \, dx \\ = \int_0^1 \alpha \frac{1}{g} \Big(\Theta^{\frac{\alpha}{n}-1} \Sigma^{\frac{n+1}{n}} - \frac{1}{1+\alpha} \Big) \Sigma V_{x\tau} \, dx.$$

Next use of the bounds (4.11) and the Poincarè inequality to obtain

$$\frac{d}{d\tau} \int_0^1 \frac{1}{g^2} V_\tau^2 dx + g(\tau) \frac{1}{C} \int_0^1 \frac{1}{g^2} V_\tau^2 dx \leq \frac{C}{g(\tau)}.$$

In turn, Gronwall's inequality implies

$$\int_0^1 V_\tau^2(x, \tau) dx \leq C, \tag{A.1}$$

and, by (4.6),

$$\int_0^1 |\Sigma_x| dx \le \left(\frac{1}{g^2} \int_0^1 V_\tau^2 dx \right)^{\frac{1}{2}} \le C e^{-\frac{1}{1+\alpha}\tau}. \tag{A.2}$$

From (4.4) we obtain the identity

$$\partial_\tau \left(e^{\frac{n-\alpha}{n(1+\alpha)}\tau} \big(\Theta^{1-\frac{\alpha}{n}}\big)_x \right) = \frac{n-\alpha}{n} e^{\frac{n-\alpha}{n(1+\alpha)}\tau} \big(\Sigma^{\frac{n+1}{n}}\big)_x,$$

which using (4.11) and (A.2) leads to using (4.11) and (A.2):

$$\int_0^1 |\Theta_x| dx \le C e^{-\frac{n-\alpha}{n(1+\alpha)}\tau}. \tag{A.3}$$

Moreover, starting from (4.5) we obtain $V_{xx} = \frac{1}{n}\Sigma^{\frac{1}{n}-1}\Theta^{\frac{\alpha}{n}}\Sigma_x + \frac{\alpha}{n}\Theta^{\frac{\alpha}{n}-1}\Theta_x\Sigma^{\frac{1}{n}}$, which is used, in conjunction with (A.2) and (A.3), to deduce

$$|V_x(x,\tau) - 1| \le \int_0^1 |V_{xx}|\, dx = O\left(e^{-\frac{n-\alpha}{n(1+\alpha)}\tau} \right). \tag{A.4}$$

Finally, from $(4.4)_2$, (4.5), and (A.4), we obtain

$$\partial_\tau \left(e^\tau \frac{1}{1+\alpha} \Theta^{1+\alpha} \right) = e^\tau V_x^{n+1} = e^\tau \left(1 + O\big(e^{-\frac{n-\alpha}{n(1+\alpha)}\tau}\big) \right)^{n+1},$$

which implies

$$\Theta = (1+\alpha)^{\frac{1}{1+\alpha}} + O\left(e^{-\frac{n-\alpha}{n(1+\alpha)}\tau} \right), \tag{A.5}$$

$$\Sigma = \Theta^{-\alpha} V_x^n = (1+\alpha)^{-\frac{\alpha}{1+\alpha}} + O\left(e^{-\frac{n-\alpha}{n(1+\alpha)}\tau} \right). \tag{A.6}$$

The estimates (A.4), (A.5), and (A.6) together with (4.3) yield the asymptotic behaviors stated in Theorem 4.1. □

Proof of Theorem 4.2. Let $q = -\alpha + n < 0$, consider initial data satisfying (4.16) and (4.17), and let (v, θ, σ) be a smooth solution of (4.1), (4.2) defined on a maximal interval of existence $[0, T_\star)$. The stress σ satisfies the boundary value problem for the parabolic equation

$$\begin{aligned} \sigma_t &= \frac{n}{r}\ \theta^{-\frac{\alpha}{n}}\ \sigma^{\frac{n-1}{n}} \sigma_{xx} - \alpha\theta^{\frac{\alpha}{n}-1}\ \sigma^{2+\frac{1}{n}}, \\ \sigma_x(0,t) &= \sigma_x(1,t) = 0. \end{aligned} \tag{A.7}$$

By the maximum principle, $\sigma(x,t) > 0$. Let $S(t)$ be the solution of the initial value problem

$$\begin{cases} \dfrac{dS}{dt} = -\alpha \underline{\theta_0}^{\frac{\alpha}{n}-1} S^{2+\frac{1}{n}}, \\ S(0) = S_0 = \sup_{x\in[0,1]} \sigma_0(x), \end{cases} \tag{A.8}$$

where $\underline{\theta_0} = \inf_{x\in[0,1]} \theta_0(x)$. S is given by the formula

$$S(t) = \left(S_0^{-\frac{n+1}{n}} + \alpha \frac{n+1}{n} \underline{\theta_0}^{\frac{\alpha}{n}-1} t \right)^{-\frac{n}{n+1}}.$$

For $\frac{\alpha}{n} > 1$, S is a supersolution of (A.7), and thus

$$0 < \sigma(x,t) \le O\left((t+1)^{-\frac{n}{n+1}}\right), \quad x \in [0,1],\ 0 < t < T_*. \tag{A.9}$$

We multiply $(4.1)_1$ by $\sigma^{\frac{1}{n}-1} v_t$, integrate over $[0,1]$ by parts, and use $(4.1)_2$ to obtain

$$\begin{aligned}\frac{n}{2}\int_0^1 \theta^{-\frac{\alpha}{n}} v_x^2 dx + \int_0^t\int_0^1 \sigma^{\frac{1-n}{n}} v_t^2 dxdt + \frac{\alpha}{2}\int_0^t\int_0^1 \theta^{-\frac{\alpha}{n}-1-\alpha} v_x^{n+3} dx \\ = \frac{n}{2}\int_0^1 \theta_0(x)^{-\frac{\alpha}{n}} v_{0x}(x)^2 dx = I_0.\end{aligned}$$

Next, we employ the calculus inequality

$$f^2(x) - \left(1+\frac{1}{r}\right) f^2(y) \le (1+r)\left(f(x)-f(y)\right)^2 \le (1+r)\int_0^1 f_x^2 d\xi$$

and put $f = \sigma^{\frac{n+1}{2n}}$ in order to obtain (for $r=2$, say)

$$\sigma^{\frac{n+1}{n}}(x,t) - \frac{3}{2}\sigma^{\frac{n+1}{n}}(y,t) \le 3\Big(\frac{n+1}{2n}\Big)^2 \int_0^1 \sigma^{\frac{1-n}{n}} v_t^2 dx.$$

Let now the data be as in (4.16), (4.17). In the region $\frac{\alpha}{n} > 1$, the identity

$$\partial_t \theta^{-\left(\frac{\alpha}{n}-1\right)} = -\left(\frac{\alpha}{n}-1\right)\sigma^{\frac{n+1}{n}},$$

taken at two distinct points x and y, gives

$$\begin{aligned}\theta^{-\left(\frac{\alpha}{n}-1\right)}(x,t) - \frac{3}{2}\theta^{-(\frac{\alpha}{n}-1)}(y,t) &= \left(\theta_0^{-\left(\frac{\alpha}{n}-1\right)}(x) - \frac{3}{2}\theta_0^{-\left(\frac{\alpha}{n}-1\right)}(y)\right) \\ &\quad - \left(\frac{\alpha}{n}-1\right)\int_0^t \left(\sigma^{\frac{n+1}{n}}(x,\tau) - \frac{3}{2}\sigma^{\frac{n+1}{n}}(y,\tau)\right) d\tau \\ &\ge \theta_0^{-\left(\frac{\alpha}{n}-1\right)}(x) - \frac{3}{2}\theta_0^{-\left(\frac{\alpha}{n}-1\right)}(y) - C(\alpha,n) I_0 \\ &= m(x,y),\end{aligned} \tag{A.10}$$

where $C(\alpha,n)$ is an explicit positive constant. Suppose now that the solution θ does not blow up. Then $\theta(x,t)$ can be estimated outside the band I_δ by the bound (A.10). The latter is, of course, meaningful only when $m(x,y) > 0$. This can be achieved provided the value $\theta_0(y) = U(y)$ in (4.17) is sufficiently large and the base state $\bar{\theta}$ suitably chosen.

Once x, y are selected so that $m(x,y) > 0$, (A.10) provides a bound for the temperature

$$\theta(x,t) \le M \quad x \notin I_\delta,\ 0 < t < \infty.$$

Since θ is increasing, it converges to a limiting profile for $x \notin I_\delta$. Inside the band θ might increase indefinitely. In addition, (A.9) and (4.2) imply

$$v_x(x,t) = O\left((t+1)^{-\frac{1}{n+1}}\right), \quad x \notin I_\delta,\ 0 < t < \infty,$$

in turn giving (4.18). □

REFERENCES

[1] L. Anand, K.H. Kim, and T.G. Shawki, *Onset of shear localization in viscoplastic solids*, J. Mech. Phys. Solids, 35 (1987), pp. 407–429.

[2] Th. Baxevanis, Th. Katsaounis and A. Tzavaras, *Adaptive finite element computations of shear band formation*, submitted, preprint available at http://www.tem.uoc.gr/∼thodoros.

[3] M. Bertsch, L.A. Peletier, and S.M. Verduyn Lunel, *The effect of temperature dependent viscosity on shear flow of incompressible fluids*, SIAM J. Math. Anal., 22 (1991), pp. 328–343.

[4] N. Charalambakis and F. Murat, *Weak solutions to the initial-boundary value problem for the shearing of non-homogeneous thermoviscoplastic materials*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 257–265.

[5] G.-Q. Chen, C.D. Levermore, and T.P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[6] K. Chueh, C. Conley, and J. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 372–411.

[7] L.S. Costin, E.E. Crisman, R.H. Hawley, and J. Duffy, *On the localization of plastic flow in mild steel tubes under dynamic torsional loading*, in Proceedings of the 2nd Conference on the Mechanical Properties of Materials at High Rates of Strain, Inst. Phys. Conf. Ser. 47, Oxford, 90, 1979.

[8] R.J. Clifton, J. Duffy, K.A. Hartley, and T.G. Shawki, *On critical conditions for shear band formation at high strain rates*, Scripta Met., 18 (1984), pp. 443–448.

[9] C.M. Dafermos and L. Hsiao, *Adiabatic shearing of incompressible fluids with temperature dependent viscosity*, Quart. Appl. Math., 41 (1983), pp. 45–58.

[10] J.A. Dilellio and W.E. Olmstead, *Shear band formation due to a thermal flux inhomogeneity*, SIAM J. Appl. Math., 57 (1997), pp. 959–971.

[11] D.J. Estep, S.M.V. Lunel, and R.D. Williams, *Analysis of shear layers in a fluid with temperature-dependent viscosity*, Comp. Phys., 173 (2001), pp. 17–60.

[12] R.P. Flemming, S.H. Davis, and W.E. Olmstead, *Shear localization with an Arrhenius flow law*, SIAM J. Appl. Math., 60 (2000), pp. 1867–1886.

[13] C. Fressengeas and A. Molinari, *Instability and localization of plastic flow in shear at high strain rates*, J. Mech. Phys. Solids, 35 (1987), pp. 185–211.

[14] K.A. Hartley, J. Duffy, and R.J. Hawley, *Measurement of the temperature profile during shear band formation in steels deforming at high-strain rates*, J. Mech. Phys. Solids, 35 (1987), pp. 283–301.

[15] R.W. Klopp. R.J. Clifton, and T.G. Shawki, *Pressure-shear Impact and the Dynamic Viscoplastic Response of Metals*, Mech. Mater. 4, Springer, New York, 1985, pp. 375–385.

[16] J.H. Maddocks and R. Malek-Madani, *Steady-state shear-bands in thermoplasticity.* I. *Vanishing yield stress*, Internat. J. Solids Structures, 29 (1992), pp. 2039–2061.

[17] A. Molinari and R.J. Clifton, *Analytical characterization of shear localization in thermoviscoplastic materials*, J. Appl. Mech., 54 (1987), pp. 806–812.

[18] A.E. Tzavaras, *Shearing of materials exhibiting thermal softening or temperature dependent viscosity*, Quart. Appl. Math., 44 (1986), pp. 1–12.

[19] A.E. Tzavaras, *Plastic shearing of materials exhibiting strain hardening or strain softening*, Arch. Ration. Mech. Anal., 94 (1986), pp. 39–58.

[20] A.E. Tzavaras, *Effect of thermal softening in shearing of strain-rate dependent materials*, Arch. Ration. Mech. Anal., 99 (1987), pp. 349–374.

[21] A.E. Tzavaras, *Strain softening in viscoelasticity of the rate type*, J. Integral Equations Appl., 3 (1991), pp. 195–238.

[22] A.E. Tzavaras, *Nonlinear analysis techniques for shear band formation at high strain rates*, Appl. Mech. Rev., 45 (1992), pp. S82–S94.

[23] A.E. Tzavaras, *Shear strain localization in plastic deformations*, in Shock Induced Transitions and Phase Structures in General Media, IMA Vol. Math. Appl. 52, J. E. Dunn, R. Fosdick, and M. Slemrod, eds., Springer, New York, 1993, pp. 231–250.

[24] T.W. Wright and J.W. Walter, *On stress collapse in adiabatic shear bands*, J. Mech. Phys. Solids, 35 (1988), pp. 701–720.

[25] T.W. Wright and H. Ockendon, *A model for fully formed shear bands*, J. Mech. Phys. Solids, 40 (1992), pp. 1217–1226.

[26] J.W. Walter, *Numerical experiments on adiabatic shear band in one space dimension*, Int. J. Plasticity, 8 (1992), pp. 657–693.

[27] C. Zener and J.H. Hollomon, *Effect of strain rate upon plastic flow of steel*, J. Appl. Phys., 15 (1944), pp. 22–32.

SIAM J. APPL. MATH.
Vol. 69, No. 6, pp. 1644–1661

A NONAUTONOMOUS JUVENILE-ADULT MODEL: WELL-POSEDNESS AND LONG-TIME BEHAVIOR VIA A COMPARISON PRINCIPLE*

AZMY S. ACKLEH† AND KENG DENG†

Abstract. A nonautonomous nonlinear continuous juvenile-adult model where juveniles and adults depend on different resources is developed. It is assumed that juveniles are structured by age, while adults are structured by size. Existence-uniqueness results are proved using the monotone method based on a comparison principle established in this paper. Conditions on the model parameters that lead to extinction or persistence of the population are obtained via the upper-lower solution technique.

Key words. continuous juvenile-adult model, comparison principle, existence-uniqueness, extinction, uniform persistence

AMS subject classifications. 92D25, 35B40, 35L50

DOI. 10.1137/080723673

1. Introduction.

Amphibians have a biphasic lifestyle where juveniles (tadpoles) live in water while adults (frogs) live on land. For many amphibians tadpoles are herbivorous, while adults are carnivorous. Thus, juveniles and adults depend on different resources, and no resource competition takes place between them. In [2, 7, 9], we developed and analyzed discrete-time discrete-stage models that describe the dynamics of such populations. A particular amphibian example that motivated these theoretical studies is the green tree frog (*Hyla cinerea*) which we have been monitoring since 2004 [26]. Similar discrete-time discrete-stage models have been formulated and studied in [14, 31]. These models have been applied to two amphibian species, *Bufo boreas* and *Ambystoma macrodactylum.* Elasticity analysis was used to determine the most influential stage survival rate on amphibian declines, a problem which was extensively discussed in [16, 29].

In this paper, we extend our modeling efforts and develop a nonautonomous continuous age-size–structured model which describes the dynamics of a population composed of juveniles and adults who depend on different resources. In our setting below we assume that juveniles are structured by age (e.g., for the green tree frogs it requires 5–6 weeks for a juvenile to metamorphose into an adult [21, 22, 28]). We assume that adults are structured by size, since fertility and mortality depend on the size of the adult (see, e.g., [33] for the green tree frogs). Furthermore, we assume that the vital parameters are time-dependent functions due to the seasonality of such populations.

Well-posedness and long-time behavior of continuous nonlinear autonomous age-size–structured population models have been investigated in many articles (e.g., see [5, 6, 11, 12, 13, 18, 19, 23, 27, 32] and the references cited therein). Structured juvenile-adult models with time-independent parameters have also been developed and studied in the literature. For example, in [15] a juvenile-adult model was developed with both juvenile and adult populations being age-structured. Therein, the

*Received by the editors May 8, 2008; accepted for publication (in revised form) November 25, 2008; published electronically March 27, 2009. This work was supported in part by the National Science Foundation under grant DMS-0718465.

http://www.siam.org/journals/siap/69-6/72367.html

†Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504 (ackleh@louisiana.edu, deng@louisiana.edu).

authors tackled the question of whether juvenile versus adult intraspecific competition is stabilizing or destabilizing. It was shown that suppressed adult fertility due to juvenile competition is destabilizing in that equilibrium levels are lowered and equilibrium resilience is weakened. However, the effect of increased juvenile mortality due to adult competition is complicated in that equilibrium levels are lowered but the resilience can be weakened or strengthened. In [20] a nonlinear size-structured juvenile-adult model was developed, and the linearized dynamical behavior of stationary solutions was analyzed using semigroup theory.

As is discussed in [14], the growth, reproduction, and mortality rates of many biological populations are subject to regular (time) fluctuations, which is the case for amphibians. Thus, in this paper we study the existence-uniqueness and long-time behavior of solutions to a nonautonomous nonlinear structured juvenile-adult model. Our arguments are different from those used in the above-mentioned articles and are based on a novel definition of upper and lower solutions, the establishment of a comparison principle, and the construction of monotone approximations. This comparison principle extends those developed in [10, 24, 25] to a system of nonlocal nonlinear first order hyperbolic equations. However, its establishment is distinct from those in [10, 24, 25], but is in the spirit of the one we developed in [4] for a size-structured model.

Other traditional approaches, including the employment of the method of characteristics to convert the problem to a system of delay integral equations and then apply the fixed point theory (e.g., [11]), may perhaps work for establishing existence-uniqueness results. However, we are unaware of abstract results for nonautonomous delay systems such as those developed in [17] for the autonomous counterpart, which can be used for investigating the long-time behavior of solutions to the model presented here.

The paper is organized as follows. In section 2 we introduce the model. In section 3 we give the definition of upper and lower solutions and establish a comparison principle. In section 4 we develop monotone sequences and prove their convergence to the unique solution of the model. In section 5 we investigate the asymptotic behavior of the model. Finally, concluding remarks are given in section 6.

2. The juvenile-adult model.

Let $J(a,t)$ and $A(x,t)$ denote the densities of juveniles of age a and adults of size x, respectively, at time t. Thus, $\int_{a_1}^{a_2} J(a,t)da$ denotes the number of juveniles in the age interval (a_1,a_2) at time t. We denote by $a_{\max}$ the age at which juveniles (tadpoles) metamorphose into adults (frogs) of minimum size $x_{\min}$, and $x_{\max}$ denotes the maximum size of adults. Hence, $\int_{x_1}^{x_2} A(x,t)dx$ denotes the number of adults in the size interval (x_1,x_2) at time t. We assume that juveniles live in an environment with abundant resources and thus do not compete, while adults live in an environment with limiting resources and thus competition between them takes place. Consider the following system of partial differential equations which describe the dynamics of interacting juveniles and adults:

$$\begin{array}{lll}
J_t(a,t) + J_a(a,t) + \nu(a,t)J(a,t) = 0, & 0 < a < a_{\max}, & 0 < t < T,\\
A_t(x,t) + (g(x,t)A(x,t))_x + \mu(x,t,\varphi(t))A(x,t) = 0, & x_{min} < x < x_{\max}, & 0 < t < T,\\
J(0,t) = \displaystyle\int_{x_{\min}}^{x_{\max}} \beta(x,t,\varphi(t))A(x,t)dx, & & 0 < t < T,\\
g(x_{\min},t)A(x_{\min},t) = J(a_{\max},t), & & 0 < t < T,\\
J(a,0) = J_0(a), & 0 \le a \le a_{\max}, & \\
A(x,0) = A_0(x), & x_{\min} \le x \le x_{\max}, &
\end{array} \tag{2.1}$$

where $\varphi(t) = \int_{x_{min}}^{x_{\max}} A(x,t)dx$ is the total population of adults. The parameters ν and μ are mortality rates for juveniles and adults, respectively. The functions g and β are the growth and reproduction rates, respectively, for adults. Since amphibians breed seasonally, the birth rate β typically depends on t and is positive during the breeding season and zero otherwise (see, e.g., [7, 9]).

Throughout the discussion, for convenience we denote $\frac{\partial \mu}{\partial \varphi}$ and $\frac{\partial \beta}{\partial \varphi}$ by μ_φ and β_φ, respectively. We assume that the parameters in (2.1) satisfy the following assumptions:

(A1) $g \in C^1([x_{\min}, x_{\max}] \times [0,T])$. Furthermore, $g(x,t) > 0$ for $(x,t) \in [x_{\min}, x_{\max}) \times [0,T]$ and $g(x_{\max}, t) = 0$ for $t \in [0,T]$.

(A2) $\nu \in L^\infty((0, a_{\max}) \times (0,T))$ is nonnegative.

(A3) $\mu(\cdot,\cdot,\varphi) \in L^\infty((x_{\min}, x_{\max}) \times (0,T))$ and for $(x,t,\varphi) \in (x_{\min}, x_{\max}) \times (0,T) \times [0,\infty)$, $\mu(x,t,\varphi)$ is nonnegative. Furthermore, μ is continuously differentiable with respect to φ with $\mu_\varphi \geq 0$.

(A4) $\beta(\cdot,\cdot,\varphi) \in L^\infty((x_{\min}, x_{\max}) \times (0,T))$ and for $(x,t,\varphi) \in (x_{\min}, x_{\max}) \times (0,T) \times [0,\infty)$, $\beta(x,t,\varphi)$ is nonnegative. Furthermore, β is continuously differentiable with respect to φ with $\beta_\varphi \leq 0$.

(A5) $J_0 \in L^\infty(0, a_{\max})$ is nonnegative.

(A6) $A_0 \in L^\infty(x_{\min}, x_{\max})$ is nonnegative.

3. Comparison principle.

We first introduce the definition of the solution of problem (2.1) via the method of characteristics. For the first equation in (2.1), the characteristic curves can be easily obtained. For the second equation in (2.1), the characteristic curves are given by

$$\begin{cases} \dfrac{d}{ds}t(s) = 1, \\ \dfrac{d}{ds}x(s) = g(x(s), t(s)). \end{cases} \tag{3.1}$$

Under assumption (A1), equation (3.1) has a unique solution for any initial point $(x(s_0), t(s_0))$. Parameterizing the characteristic curves with the variable t, a characteristic curve passing through $(\hat{x}, \hat{t})$ is given by $(X(t;\hat{x},\hat{t}), t)$, where X satisfies

$$\frac{d}{dt}X(t;\hat{x},\hat{t}) = g(X(t;\hat{x},\hat{t}), t)$$

and $X(\hat{t};\hat{x},\hat{t}) = \hat{x}$. By (A1), the function X is strictly increasing, and therefore a unique inverse function $\Gamma(x;\hat{x},\hat{t})$ exists. Let $G(x) = \Gamma(x; x_{\min}, 0)$; then $(x, G(x))$ represents the characteristic curve passing through $(x_{\min}, 0)$, and this curve divides the (x,t)-plane into two parts. Hence, we can define the solution of problem (2.1) as follows:

$$J(a,t) = J_0(a-t)\exp\left(-\int_0^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t \leq a, \tag{3.2}$$

$$J(a,t) = \int_{x_{\min}}^{x_{\max}} \beta(x, t-a, \varphi(t-a))A(x,t-a)dx \exp\left(-\int_{t-a}^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t > a, \tag{3.3}$$

$$A(x,t) = A_0(X(0;x,t)) \exp\left\{-\int_0^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,\varphi(\tau))]d\tau\right\} \quad \text{if } t \le G(x), \tag{3.4}$$

$$A(x,t) = \frac{J(a_{\max},\Gamma(x_{\min};x,t))}{g(x_{\min},\Gamma(x_{\min};x,t))} \exp\left\{-\int_{\Gamma(x_{\min};x,t)}^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,\varphi(\tau))]d\tau\right\} \quad \text{if } t > G(x). \tag{3.5}$$

We then introduce the definition of a pair of coupled upper and lower solutions of problem (2.1).

Definition 3.1. *A pair of functions* ($\overline{J}(a,t), \overline{A}(x,t)$) *and* ($\underline{J}(a,t), \underline{A}(x,t)$) *are called an upper solution and a lower solution, respectively, of* (2.1) *if the following statements hold:*

(i) $\overline{J}, \underline{J} \in L^\infty((0,a_{\max}) \times (0,T))$ *and* $\overline{A}, \underline{A} \in L^\infty((x_{\min},x_{\max}) \times (0,T))$.

(ii)

$$\overline{J}(a,t) \ge J_0(a-t) \exp\left(-\int_0^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t \le a. \tag{3.6}$$

$$\overline{J}(a,t) \ge \int_{x_{\min}}^{x_{\max}} \beta(x,t-a,\underline{\varphi}(t-a))\overline{A}(x,t-a)dx \exp\left(-\int_{t-a}^t \nu(a-t+\tau,\tau)d\tau\right) \quad \text{if } t > a. \tag{3.7}$$

(iii) $\overline{A}(x,0) \ge A_0(x) \ge \underline{A}(x,0)$ *a.e. in* $(x_{\min},x_{\max})$. *For each* $t \in (0,T)$ *and every nonnegative* $\xi \in C^1([x_{\min},x_{\max}] \times [0,T])$,

$$\begin{aligned} \int_{x_{\min}}^{x_{\max}} \overline{A}(x,t)\xi(x,t)dx \ge & \int_{x_{\min}}^{x_{\max}} \overline{A}(x,0)\xi(x,0)dx + \int_0^t \overline{J}(a_{\max},\tau)\xi(x_{\min},\tau)d\tau \\ & + \int_0^t \int_{x_{\min}}^{x_{\max}} [\xi_\tau(x,\tau) + g(x,\tau)\xi_x(x,\tau)]\overline{A}(x,\tau)dxd\tau \\ & - \int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\underline{\varphi}(\tau))\overline{A}(x,\tau)\xi(x,\tau)dxd\tau. \end{aligned} \tag{3.8}$$

($\underline{J}, \underline{A}$) *satisfies* (3.6)–(3.8), *respectively, by replacing " $\ge$ " with " $\le$ " and by interchanging* $\overline{J}$ *with* $\underline{J}$, $\overline{A}$ *with* $\underline{A}$, *and* $\underline{\varphi}$ *with* $\overline{\varphi}$, *where* $\underline{\varphi}(t) = \int_{x_{\min}}^{x_{\max}} \underline{A}(x,t)dx$ *and* $\overline{\varphi}(t) = \int_{x_{\min}}^{x_{\max}} \overline{A}(x,t)dx$.

Remark 3.2. Inequality (3.8), which is used only for the establishment of the comparison principle, is motivated by an alternative definition of a weak solution of the second equation in (2.1) given by

$$\begin{aligned} \int_{x_{\min}}^{x_{\max}} A(x,t)\xi(x,t)dx = & \int_{x_{\min}}^{x_{\max}} A(x,0)\xi(x,0)dx + \int_0^t J(a_{\max},\tau)\xi(x_{\min},\tau)d\tau \\ & + \int_0^t \int_{x_{\min}}^{x_{\max}} [\xi_\tau(x,\tau) + g(x,\tau)\xi_x(x,\tau)]A(x,\tau)dxd\tau \\ & - \int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\varphi(\tau))A(x,\tau)\xi(x,\tau)dxd\tau \end{aligned} \tag{3.9}$$

for each $t \in (0,T)$ and every $\xi \in C^1([x_{\min}, x_{\max}] \times [0,T])$. The solution (3.9) is derived by formally multiplying the second equation of (2.1) by a test function ξ and integrating by parts (cf. [1, 8]).

Remark 3.3. For the *linear* problem

$$\begin{aligned}
&A_t(x,t) + (g(x,t)A(x,t))_x + \mu(x,t)A(x,t) = 0, \quad x_{min} < x < x_{\max}, \quad 0 < t < T,\\
&g(x_{\min},t)A(x_{\min},t) = h(t), \qquad\qquad\qquad\qquad\qquad\qquad 0 < t < T,\\
&A(x,0) = A_0(x), \qquad\qquad\qquad\qquad\qquad x_{\min} \le x \le x_{\max},
\end{aligned} \tag{3.10}$$

the solution satisfies both the mild form given by (3.4)–(3.5) and the weak form given by (3.9) with $J(a_{\max},t)$ replaced by $h(t)$. Moreover, the two sequences that we construct in section 4 and apply the comparison principle to are linear and of the form (3.10).

Based on such a definition, we can establish the following comparison result.

THEOREM 3.4. *Suppose that* (A1)–(A6) *hold. Let* $(\overline{J}, \overline{A})$ *and* $(\underline{J}, \underline{A})$ *be a nonnegative upper solution and a nonnegative lower solution, respectively, of* (2.1)*. Then* $\overline{J} \ge \underline{J}$ *a.e. in* $(0, a_{\max}) \times (0, T_0)$ *and* $\overline{A} \ge \underline{A}$ *a.e. in* $(x_{\min}, x_{\max}) \times (0, T_0))$*, where* $T_0 = \min\{T, a_{\max}\}$*.*

Proof. Let $K = \underline{J} - \overline{J}$ and $B = \underline{A} - \overline{A}$. In view of (3.6), $K(a,t) \le 0$ for $t \le a$. In particular, $K(a_{\max},t) \le 0$. Furthermore, B satisfies

$$B(x,0) = \underline{A}(x,0) - \overline{A}(x,0) \le 0 \quad \text{a.e. in } (x_{\min}, x_{\max}) \tag{3.11}$$

and

$$\begin{aligned}
\int_{x_{\min}}^{x_{\max}} B(x,t)\xi(x,t)dx \le& \int_{x_{\min}}^{x_{\max}} B(x,0)\xi(x,0)dx + \int_0^t K(a_{\max},\tau)\xi(x_{\min},\tau)d\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} [\xi_\tau(x,\tau) + g(x,\tau)\xi_x(x,\tau)]B(x,\tau)dxd\tau\\
&-\int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\overline{\varphi}(\tau))B(x,\tau)\xi(x,\tau)dxd\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} \xi(x,\tau)C_1(x,\tau)\int_{x_{\min}}^{x_{\max}} B(y,\tau)dydxd\tau,
\end{aligned} \tag{3.12}$$

where $C_1(x,t) = \overline{A}(x,t)\mu_\varphi(x,t,\theta_1(t))$ with $\theta_1(t)$ between $\overline{\varphi}(t)$ and $\underline{\varphi}(t)$.

Let $\xi(x,t) = e^{\lambda t}\zeta(x,t)$, where $\zeta \in C^1([x_{\min}, x_{\max}] \times [0,T_0])$ and λ (> 0) is chosen so that $\lambda - \mu(x,t,\overline{\varphi}(t)) \ge 0$ on $(x_{\min}, x_{\max}) \times (0,T_0)$. Then we find

$$\begin{aligned}
e^{\lambda t}\int_{x_{\min}}^{x_{\max}} B(x,t)\zeta(x,t)dx \le& \int_{x_{\min}}^{x_{\max}} B(x,0)\zeta(x,0)dx + \int_0^t e^{\lambda\tau} K(a_{\max},\tau)\zeta(x_{\min},\tau)d\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} e^{\lambda\tau}[\zeta_\tau(x,\tau) + g(x,\tau)\zeta_x(x,\tau)]B(x,\tau)dxd\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} e^{\lambda\tau}[\lambda - \mu(x,\tau,\overline{\varphi}(\tau))]B(x,\tau)\zeta(x,\tau)dxd\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} e^{\lambda\tau}\zeta(x,\tau)C_1(x,\tau)\int_{x_{\min}}^{x_{\max}} B(y,\tau)dydxd\tau.
\end{aligned} \tag{3.13}$$

We now set up a backward problem as follows:

$$\begin{aligned}
&\zeta_\tau + g\zeta_x = 0, \qquad x_{\min} < x < x_{\max}, \quad 0 < \tau < t,\\
&\zeta(x_{\max}, \tau) = 0, \qquad\qquad\qquad\qquad\quad 0 < \tau < t,\\
&\zeta(x, t) = \chi(x), \qquad x_{\min} < x < x_{\max}.
\end{aligned}$$

Here $\chi \in C_0^\infty(x_{\min}, x_{\max})$ with $0 \le \chi \le 1$. The above problem can be solved by the method of characteristics, and the solution satisfies $0 \le \zeta(x,t) \le 1$ on $[x_{\min}, x_{\max}] \times [0, T_0]$. Since $B(x,0) \le 0$ and $K(a_{\max}, t) \le 0$, we find

$$\begin{aligned}
e^{\lambda t}\int_{x_{\min}}^{x_{\max}} B(x,t)\chi(x)dx \le{}& \int_0^t \int_{x_{\min}}^{x_{\max}} e^{\lambda\tau}[\lambda - \mu(x,\tau,\overline{\varphi}(\tau))]B(x,\tau)\zeta(x,\tau)dxd\tau\\
&+\int_0^t \int_{x_{\min}}^{x_{\max}} e^{\lambda\tau}\zeta(x,\tau)C_1(x,\tau)\int_{x_{\min}}^{x_{\max}} B(y,\tau)dydxd\tau.
\end{aligned} \tag{3.14}$$

Therefore, we have

$$\int_{x_{\min}}^{x_{\max}} B(x,t)\chi(x)dx \le c_1 \int_0^t \int_{x_{\min}}^{x_{\max}} B^+(x,\tau)dxd\tau, \tag{3.15}$$

where

$$c_1 = \sup_{(x,t)\in[x_{\min},x_{\max}]\times[0,T_0]} \left[\lambda - \mu(x,t,\overline{\varphi}(t)) + \int_{x_{\min}}^{x_{\max}} C_1(x,t)dx\right]$$

and $B^+(x,t) = \max\{0, B(x,t)\}$.

Since this inequality holds for every χ, we can now choose a sequence $\{\chi_n\}$ on $(x_{\min}, x_{\max})$ converging a.e. to

$$\chi(x) = \begin{cases} 1 & \text{if } B(x,t) > 0,\\ 0 & \text{otherwise.}\end{cases}$$

Consequently, we find

$$\int_{x_{\min}}^{x_{\max}} B^+(x,t)dx \le c_1 \int_0^t \int_{x_{\min}}^{x_{\max}} B^+(x,\tau)dxd\tau,$$

which by Gronwall's inequality leads to

$$\int_{x_{\min}}^{x_{\max}} B^+(x,t)dx = 0,$$

i.e., $B(x,t) \le 0$. Then it follows from (3.7) that for $t > a$

$$\begin{aligned}
K(a,t) \le{}& \int_{x_{\min}}^{x_{\max}} \beta(x, t-a, \overline{\varphi}(t-a))B(x,t-a)dx \exp\left(-\int_{t-a}^t \nu(a-t+\tau,\tau)d\tau\right)\\
&-\int_{x_{\min}}^{x_{\max}} D(x,t-a)\int_{x_{\min}}^{x_{\max}} B(y,t-a)dydx \exp\left(-\int_{t-a}^t \nu(a-t+\tau,\tau)d\tau\right)\\
\le{}& 0,
\end{aligned} \tag{3.16}$$

where $D(x,t) = \overline{A}(x,t)\beta_\varphi(x,t,\theta_2(t))$ with $\theta_2(t)$ between $\overline{\varphi}(t)$ and $\underline{\varphi}(t)$. This completes the proof. □

Based on the definition (3.2)–(3.5), we can also establish the following uniqueness result.

THEOREM 3.5. *Let* $(J(a,t), A(x,t))$ *be a nonnegative solution of problem* (2.1) *for* $0 \le t \le T_0$. *Then* (J, A) *is unique.*

Proof. Suppose that $(J_1(a,t), A_1(x,t))$ and $(J_2(a,t), A_2(x,t))$ are two nonnegative solutions of (2.1). Clearly, $J_1(a,t) = J_2(a,t)$ for $t \le a$. By (3.4) we have

$$A_1(x,t) - A_2(x,t) = A_0(X(0;x,t)) \exp\left(-\int_0^t g_x(X(\tau;x,t),\tau)d\tau - \theta_3(x,t)\right) \times \int_0^t C_2(x,\tau)\int_{x_{\min}}^{x_{\max}} [A_1(x,\tau) - A_2(x,\tau)]dxd\tau,$$

where $\theta_3(x,t)$ is between $\int_0^t \mu(X(\tau;x,t),\tau,\varphi_1(\tau))d\tau$ and $\int_0^t \mu(X(\tau;x,t),\tau,\varphi_2(\tau))d\tau$ with $\varphi_1(\tau) = \int_{x_{\min}}^{x_{\max}} A_1(x,\tau)dx$ and $\varphi_2(\tau) = \int_{x_{\min}}^{x_{\max}} A_2(x,\tau)dx$, and $C_2(x,\tau) = \mu_\varphi(X(\tau;x,t),\tau,\theta_4(\tau))$ with $\theta_4(\tau)$ between $\varphi_1(\tau)$ and $\varphi_2(\tau)$. Thus, for $t \le G(x)$

$$|A_1(x,t) - A_2(x,t)| \le c_2 \int_0^t \int_{x_{\min}}^{x_{\max}} |A_1(x,\tau) - A_2(x,\tau)|dxd\tau. \tag{3.17}$$

On the other hand, since $J_1(a_{\max}, \Gamma(x_{\min};x,t)) = J_2(a_{\max}, \Gamma(x_{\min};x,t))$, by (3.5) we have

$$\begin{aligned} &A_1(x,t) - A_2(x,t) \\ &\quad = \frac{J(a_{\max},\Gamma(x_{\min};x,t))}{g(x_{\min},\Gamma(x_{\min};x,t))} \exp\left(-\int_{\Gamma(x_{\min};x,t)}^t g_x(X(\tau;x,t),\tau)d\tau - \theta_5(x,t)\right) \\ &\qquad \times \int_{\Gamma(x_{\min};x,t)}^t C_3(x,\tau)\int_{x_{\min}}^{x_{\max}} [A_1(x,\tau) - A_2(x,\tau)]dxd\tau, \end{aligned}$$

where $\theta_5(x,t)$ is between $\int_{\Gamma(x_{\min};x,t)}^t \mu(X(\tau;x,t),\tau,\varphi_1(\tau))d\tau$ and $\int_{\Gamma(x_{\min};x,t)}^t \mu(X(\tau;x,t),\tau,\varphi_2(\tau))d\tau$, and $C_3(x,\tau) = \mu_\varphi(X(\tau;x,t),\tau,\theta_6(\tau))$ with $\theta_6(\tau)$ between $\varphi_1(\tau)$ and $\varphi_2(\tau)$. Thus, for $t > G(x)$

$$\begin{aligned} |A_1(x,t) - A_2(x,t)| &\le c_3 \int_{\Gamma(x_{\min};x,t)}^t \int_{x_{\min}}^{x_{\max}} |A_1(x,\tau) - A_2(x,\tau)|dxd\tau \\ &\le c_3 \int_0^t \int_{x_{\min}}^{x_{\max}} |A_1(x,\tau) - A_2(x,\tau)|dxd\tau. \end{aligned} \tag{3.18}$$

A combination of (3.16) and (3.17) then yields that for any $(x,t) \in [x_{\min}, x_{\max}] \times [0, T_0]$

$$|A_1(x,t) - A_2(x,t)| \le c_4 \int_0^t \int_{x_{\min}}^{x_{\max}} |A_1(x,\tau) - A_2(x,\tau)|dxd\tau. \tag{3.19}$$

Integration of (3.19) over $(x_{\min}, x_{\max})$ gives

$$\int_{x_{\min}}^{x_{\max}} |A_1(x,t) - A_2(x,t)|dx \le c_4(x_{\max} - x_{\min}) \int_0^t \int_{x_{\min}}^{x_{\max}} |A_1(x,\tau) - A_2(x,\tau)|dxd\tau,$$

which by Gronwall's inequality implies

$$\int_{x_{\min}}^{x_{\max}} |A_1(x,t) - A_2(x,t)| dx = 0.$$

Thus, by (3.19) we have $A_1(x,t) = A_2(x,t)$ for $(x,t) \in [x_{\min}, x_{\max}] \times [0, T_0]$. Consequently, taking note of (3.3), $J_1(a,t) = J_2(a,t)$ for $a < t \le T_0$. □

4. Monotone sequences and existence of the solution. We begin with the introduction of a pair of nonnegative lower and upper solutions of problem (2.1). Let $\underline{J}^0(a,t) = 0$ and $\underline{A}^0(x,t) = 0$. For $0 \le t \le T_0$ $(\equiv a_{\max})$, we set

$$\begin{aligned}
&\overline{J}^0_t(a,t) + \overline{J}^0_a(a,t) + \nu(a,t)\overline{J}^0(a,t) = 0, && 0 < a < a_{\max}, && 0 < t < T_0,\\
&\overline{A}^0_t(x,t) + (g(x,t)\overline{A}^0(x,t))_x + \mu(x,t,0)\overline{A}^0(x,t) = 0, && x_{\min} < x < x_{\max}, && 0 < t < T_0,\\
&\overline{J}^0(0,t) = \int_{x_{\min}}^{x_{\max}} \beta(x,t,0)\overline{A}^0(x,t)dx, && && 0 < t < T_0,\\
&g(x_{\min},t)\overline{A}^0(x_{\min},t) = \overline{J}^0(a_{\max},t), && && 0 < t < T_0,\\
&\overline{J}^0(a,0) = J_0(a), && 0 \le a \le a_{\max}, &&\\
&\overline{A}^0(x,0) = A_0(x), && x_{\min} \le x \le x_{\max}. &&
\end{aligned} \tag{4.1}$$

Clearly, by the method of characteristics, $\overline{J}^0$ and $\overline{A}^0$ satisfy

$$\overline{J}^0(a,t) = J_0(a-t)\exp\left(-\int_0^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t \le a, \tag{4.2}$$

$$\overline{J}^0(a,t) = \int_{x_{\min}}^{x_{\max}} \beta(x,t-a,0)\overline{A}^0(x,t-a)dx \exp\left(-\int_{t-a}^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t > a, \tag{4.3}$$

$$\overline{A}^0(x,t) = A_0(X(0;x,t))\exp\left\{-\int_0^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,0)]d\tau\right\} \qquad \text{if } t \le G(x), \tag{4.4}$$

$$\begin{aligned}\overline{A}^0(x,t) &= \frac{\overline{J}^0(a_{\max},\Gamma(x_{\min};x,t))}{g(x_{\min},\Gamma(x_{\min};x,t))}\exp\left\{-\int_{\Gamma(x_{\min};x,t)}^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,0)]d\tau\right\}\\ &\qquad \text{if } t > G(x).\end{aligned} \tag{4.5}$$

The above representations are uncoupled, since $\overline{A}^0(x,t)$ in (4.5) is constructed from $\overline{J}^0(a,t)$ in (4.2), and then $\overline{J}^0(a,t)$ in (4.3) is obtained from (4.4) and (4.5). Therefore,

$\overline{A}^0$ satisfies

$$\begin{aligned}\int_{x_{\min}}^{x_{\max}}\overline{A}^0(x,t)\xi(x,t)dx &= \int_{x_{\min}}^{x_{\max}}\overline{A}^0(x,0)\xi(x,0)dx + \int_0^t \overline{J}^0(a_{\max},\tau)\xi(x_{\min},\tau)d\tau \\ &\quad + \int_0^t\int_{x_{\min}}^{x_{\max}}[\xi_\tau(x,\tau)+g(x,\tau)\xi_x(x,\tau)]\overline{A}^0(x,\tau)dxd\tau \\ &\quad - \int_0^t\int_{x_{\min}}^{x_{\max}}\mu(x,\tau,0)\overline{A}^0(x,\tau)\xi(x,\tau)dxd\tau\end{aligned}$$

for each $t\in(0,T)$ and every $\xi\in C^1([x_{\min},x_{\max}]\times[0,T])$, since $\overline{J}^0(a_{\max},t)$ is explicitly obtained from (4.2) (cf. [8]).

Thus, it easily follows that $(\underline{J}^0,\underline{A}^0)$ and $(\overline{J}^0,\overline{A}^0)$ are a pair of lower and upper solutions of (2.1) for $0\le t\le T_0$.

We then define two sequences $\{\underline{J}^k,\underline{A}^k\}_{k=0}^\infty$ and $\{\overline{J}^k,\overline{A}^k\}_{k=0}^\infty$ as follows: For $k=1,2,\ldots,$

$$\begin{aligned}&\underline{J}_t^k(a,t)+\underline{J}_a^k(a,t)+\nu(a,t)\underline{J}^k(a,t)=0, && 0<a<a_{\max}, \quad 0<t<T_0,\\ &\underline{A}_t^k(x,t)+(g(x,t)\underline{A}^k(x,t))_x \\ &\qquad +\mu(x,t,\overline{\varphi}^{k-1}(t))\underline{A}^k(x,t)=0, && x_{\min}<x<x_{\max}, \quad 0<t<T_0,\\ &\underline{J}^k(0,t)=\int_{x_{\min}}^{x_{\max}}\beta(x,t,\overline{\varphi}^{k-1}(t))\underline{A}^k(x,t)dx, && 0<t<T_0,\\ &g(x_{\min},t)\underline{A}^k(x_{\min},t)=\underline{J}^k(a_{\max},t), && 0<t<T_0,\\ &\underline{J}^k(a,0)=J_0(a), && 0\le a\le a_{\max},\\ &\underline{A}^k(x,0)=A_0(x), && x_{\min}\le x\le x_{\max},\end{aligned}\tag{4.6}$$

where $\overline{\varphi}^{k-1}(t)=\int_{x_{\min}}^{x_{\max}}\overline{A}^{k-1}(x,t)dx$ and

$$\begin{aligned}&\overline{J}_t^k(a,t)+\overline{J}_a^k(a,t)+\nu(a,t)\overline{J}^k(a,t)=0, && 0<a<a_{\max}, \quad 0<t<T_0,\\ &\overline{A}_t^k(x,t)+(g(x,t)\overline{A}^k(x,t))_x \\ &\qquad +\mu(x,t,\underline{\varphi}^{k-1}(t))\overline{A}^k(x,t)=0, && x_{\min}<x<x_{\max}, \quad 0<t<T_0,\\ &\overline{J}^k(0,t)=\int_{x_{\min}}^{x_{\max}}\beta(x,t,\underline{\varphi}^{k-1}(t))\overline{A}^k(x,t)dx, && 0<t<T_0,\\ &g(x_{\min},t)\overline{A}^k(x_{\min},t)=\overline{J}^k(a_{\max},t), && 0<t<T_0,\\ &\overline{J}^k(a,0)=J_0(a), && 0\le a\le a_{\max},\\ &\overline{A}^k(x,0)=A_0(x), && x_{\min}\le x\le x_{\max},\end{aligned}\tag{4.7}$$

where $\underline{\varphi}^{k-1}(t)=\int_{x_{\min}}^{x_{\max}}\underline{A}^{k-1}(x,t)dx$. Using the method of characteristics, solutions to the above problems can be found explicitly as follows:

$$\underline{J}^k(a,t)=J_0(a-t)\exp\left(-\int_0^t\nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t\le a,\tag{4.8}$$

$$\underline{J}^k(a,t) = \int_{x_{\min}}^{x_{\max}} \beta(x, t-a, \overline{\varphi}^{k-1}(t-a))\underline{A}^k(x,t-a)dx \exp\left(-\int_{t-a}^{t} \nu(a-t+\tau,\tau)d\tau\right) \quad \text{if } t > a, \tag{4.9}$$

$$\underline{A}^k(x,t) = A_0(X(0;x,t)) \exp\left\{-\int_0^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,\overline{\varphi}^{k-1}(\tau))]d\tau\right\} \quad \text{if } t \le G(x), \tag{4.10}$$

$$\begin{aligned}\underline{A}^k(x,t) = {} & \frac{\underline{J}^k(a_{\max}, \Gamma(x_{\min};x,t))}{g(x_{\min}, \Gamma(x_{\min};x,t))} \\ & \times \exp\left\{-\int_{\Gamma(x_{\min};x,t)}^{t} [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,\overline{\varphi}^{k-1}(\tau))]d\tau\right\} \\ & \text{if } t > G(x).\end{aligned} \tag{4.11}$$

And $\underline{A}^k$ also satisfies

$$\begin{aligned}\int_{x_{\min}}^{x_{\max}} \underline{A}^k(x,t)\xi(x,t)dx = {} & \int_{x_{\min}}^{x_{\max}} \underline{A}^k(x,0)\xi(x,0)dx + \int_0^t \underline{J}^k(a_{\max},\tau)\xi(x_{\min},\tau)d\tau \\ & + \int_0^t \int_{x_{\min}}^{x_{\max}} [\xi_\tau(x,\tau) + g(x,\tau)\xi_x(x,\tau)]\underline{A}^k(x,\tau)dxd\tau \\ & - \int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\overline{\varphi}^{k-1}(t))\underline{A}^k(x,\tau)\xi(x,\tau)dxd\tau\end{aligned} \tag{4.12}$$

for each $t \in (0,T)$ and every $\xi \in C^1([x_{\min}, x_{\max}] \times [0,T])$.

A representation similar to (4.8)–(4.12) can be obtained for the solution $(\overline{J}^k, \overline{A}^k)$ by interchanging $\underline{J}^k$ with $\overline{J}^k$, $\underline{A}^k$ with $\overline{A}^k$, and $\overline{\varphi}^{k-1}$ with $\underline{\varphi}^{k-1}$.

Remark 4.1. By discretizing (4.8)–(4.11) and the corresponding equations for $(\overline{J}^k, \overline{A}^k)$, one can derive a numerical scheme for solving problem (2.1).

Clearly, $\underline{J}^0 \le \underline{J}^1$ and $\underline{A}^0 \le \underline{A}^1$. Meanwhile, $\overline{J}^1 = \overline{J}^0$ for $t \le a$, which implies $\overline{A}^1 = \overline{A}^0$. Thus, $\overline{J}^1 = \overline{J}^0$ for $t > a$. Moreover, since $-\mu(x,t,\underline{\varphi}^0(t)) \ge -\mu(x,t,\underline{\varphi}^1(t))$, $\overline{A}^1(x,t)$ satisfies the following: for each $t \in (0,T_0)$ and every nonnegative $\xi \in C^1([x_{\min}, x_{\max}] \times [0,T_0]$,

$$\begin{aligned}\int_{x_{\min}}^{x_{\max}} \overline{A}^1(x,t)\xi(x,t)dx \ge {} & \int_{x_{\min}}^{x_{\max}} \overline{A}^1(x,0)\xi(x,0)dx + \int_0^t \overline{J}^1(a_{\max},\tau)\xi(x_{\min},\tau)d\tau \\ & + \int_0^t \int_{x_{\min}}^{x_{\max}} [\xi_\tau(x,\tau) + g(x,\tau)\xi_x(x,\tau)]\overline{A}^1(x,\tau)dxd\tau \\ & - \int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\underline{\varphi}^1(\tau))\overline{A}^1(x,\tau)\xi(x,\tau)dxd\tau.\end{aligned} \tag{4.13}$$

On the other hand, since $-\mu(x,t,\overline{\varphi}^0(t)) = -\mu(x,t,\overline{\varphi}^1(t))$, $\underline{A}^1(x,t)$ satisfies (4.13) with " $\ge$ " replaced by " $\le$," $\overline{J}^1$ by $\underline{J}^1$, and $\underline{\varphi}^1$ by $\overline{\varphi}^1$. Furthermore, since $\beta(x,t,\underline{\varphi}^0(t)) \ge$

$\beta(x,t,\underline{\varphi}^1(t))$, $\overline{J}^1(a,t)$ satisfies

$$\overline{J}^1(a,t) \geq \int_{x_{\min}}^{x_{\max}} \beta(x,t-a,\underline{\varphi}^1(t-a))\overline{A}^1(x,t-a)dx \exp\left(-\int_{t-a}^{t} \nu(a-t+\tau,\tau)d\tau\right) \quad \text{if } t > a, \tag{4.14}$$

and since $\beta(x,t,\overline{\varphi}^0(t)) = \beta(x,t,\overline{\varphi}^1(t))$, $\underline{J}^1(a,t)$ satisfies (4.14) with " $\geq$ " replaced by " $\leq$," $\overline{A}^1$ by $\underline{A}^1$, and $\overline{\varphi}^1$ by $\underline{\varphi}^1$. Therefore, $(\underline{J}^1, \underline{A}^1)$ and $(\overline{J}^1, \overline{A}^1)$ are a lower solution and an upper solution, respectively, and hence $\underline{J}^1 \leq \overline{J}^1$, $\underline{A}^1 \leq \overline{A}^1$.

Assume that for some $k > 1$, $(\underline{J}^k, \underline{A}^k)$ and $(\overline{J}^k, \overline{A}^k)$ are a lower solution and an upper solution, respectively, of problem (2.1). By similar reasoning, we can show that $\underline{J}^k \leq \underline{J}^{k+1} \leq \overline{J}^{k+1} \leq \overline{J}^k$, $\underline{A}^k \leq \underline{A}^{k+1} \leq \overline{A}^{k+1} \leq \overline{A}^k$, and that $(\underline{J}^{k+1}, \underline{A}^{k+1})$ and $(\overline{J}^{k+1}, \overline{A}^{k+1})$ are also a lower solution and an upper solution, respectively, of (2.1). Thus, by induction, we obtain two monotone sequences that satisfy

$$\begin{aligned} &\underline{J}^0 \leq \underline{J}^1 \leq \cdots \leq \underline{J}^k \leq \overline{J}^k \leq \cdots \leq \overline{J}^1 \leq \overline{J}^0 \qquad && \text{a.e. in } [0,a_{\max}] \times [0,T_0], \\ &\underline{A}^0 \leq \underline{A}^1 \leq \cdots \leq \underline{A}^k \leq \overline{A}^k \leq \cdots \leq \overline{A}^1 \leq \overline{A}^0 \qquad && \text{a.e. in } [x_{\min},x_{\max}] \times [0,T_0] \end{aligned}$$

for each $k = 0,1,2,\ldots$. Hence, it follows from the monotonicity of the sequences $\{\underline{J}^k, \underline{A}^k\}$ and $\{\overline{J}^k, \overline{A}^k\}$ that there exist functions $(\underline{J}, \underline{A})$ and $(\overline{J}, \overline{A})$ such that $\underline{J}^k \to \underline{J}$ and $\overline{J}^k \to \overline{J}$ pointwise in $(0,a_{\max}) \times (0,T_0)$, and $\underline{A}^k \to \underline{A}$ and $\overline{A}^k \to \overline{A}$ pointwise in $(x_{\min},x_{\max}) \times (0,T_0)$. Clearly $\underline{J} \leq \overline{J}$ and $\underline{A} \leq \overline{A}$ a.e.

Upon establishing the monotonicity of our sequences, we can prove the following convergence result.

THEOREM 4.2. *Suppose that* (A1)–(A6) *hold. Then* $\{\underline{J}^k, \underline{A}^k\}_{k=0}^\infty$ *and* $\{\overline{J}^k, \overline{A}^k\}_{k=0}^\infty$ *converge to the unique solution* (J,A) *of problem* (2.1) *in* $L^\infty((0,a_{\max}) \times (0,T_0)) \times L^\infty((x_{\min},x_{\max}) \times (0,T_0))$.

Proof. Since the monotone sequences are bounded by $(\underline{J}^0, \underline{A}^0)$ and $(\overline{J}^0, \overline{A}^0)$, from the pointwise convergence of the sequence, the solution representation for $(\underline{J}^k, \underline{A}^k)$ given in (4.8)–(4.11), and the solution representation for $(\overline{J}^k, \overline{A}^k)$, we find that $\{\underline{J}^k, \underline{A}^k\}_{k=0}^\infty$ converges to $(\underline{J}, \underline{A})$ and $\{\overline{J}^k, \overline{A}^k\}_{k=0}^\infty$ converges to $(\overline{J}, \overline{A})$ monotonically. Here, $(\underline{J}, \underline{A})$ satisfies

$$\underline{J}(a,t) = J_0(a-t)\exp\left(-\int_0^t \nu(a-t+\tau,\tau)d\tau\right) \qquad \text{if } t \leq a, \tag{4.15}$$

$$\underline{J}(a,t) = \int_{x_{\min}}^{x_{\max}} \beta(x,t-a,\overline{\varphi}(t-a))\underline{A}(x,t-a)dx \exp\left(-\int_{t-a}^{t} \nu(a-t+\tau,\tau)d\tau\right) \quad \text{if } t > a, \tag{4.16}$$

$$\underline{A}(x,t) = A_0(X(0;x,t))\exp\left\{-\int_0^t [g_x(X(\tau;x,t),\tau) + \mu(X(\tau;x,t),\tau,\overline{\varphi}(\tau))]d\tau\right\} \quad \text{if } t \leq G(x), \tag{4.17}$$

$$\underline{A}(x,t)$$
$$= \frac{\underline{J}(a_{\max}, \Gamma(x_{\min}; x,t))}{g(x_{\min}, \Gamma(x_{\min}; x,t))} \exp\left\{ -\int_{\Gamma(x_{\min};x,t)}^{t} [g_x(X(\tau; x,t),\tau) + \mu(X(\tau;x,t),\tau,\overline{\varphi}(\tau))]d\tau \right\}$$
$$\text{if } t > G(x), \tag{4.18}$$

and $(\overline{J}, \overline{A})$ satisfies (4.15)–(4.18) by interchanging $\underline{J}$ with $\overline{J}$, $\underline{A}$ with $\overline{A}$, and $\overline{\varphi}$ with $\underline{\varphi}$.

We now show that $(\underline{J}, \underline{A}) = (\overline{J}, \overline{A})$. In view of (4.15)–(4.18), it suffices to show that $\underline{A} = \overline{A}$. To this end, let $B = \overline{A} - \underline{A}$. Since $\overline{A} \geq \underline{A}$, $B(x,t) \geq 0$ and $B(x,0) = 0$. Taking note of the fact that $\underline{J}(a_{\max}, t) = \overline{J}(a_{\max}, t)$ and choosing $\xi(x,t) \equiv 1$, we have that

$$\begin{aligned}\int_{x_{\min}}^{x_{\max}} B(x,t)dx &= -\int_0^t \int_{x_{\min}}^{x_{\max}} \mu(x,\tau,\underline{\varphi}(\tau))B(x,\tau)dxd\tau \\ &\quad + \int_0^t \int_{x_{\min}}^{x_{\max}} C_4(x,\tau) \int_{x_{\min}}^{x_{\max}} B(y,\tau)dydxd\tau \\ &\leq c_5 \int_0^t \int_{x_{\min}}^{x_{\max}} B(x,\tau)dxd\tau,\end{aligned}$$

where $C_4(x,t) = \underline{A}(x,t)\mu_\varphi(x,t,\theta_7(t))$ with $\theta_7(t)$ between $\underline{\varphi}(t)$ and $\overline{\varphi}(t)$, and $c_5 = \sup_{t\in[0,T_0]} \int_{x_{\min}}^{x_{\max}} C_4(x,t)dx$. Thus, it follows from Gronwall's inequality that $B(x,t) = 0$, i.e., $\underline{A} = \overline{A}$. Defining the common limit by (J, A), we find that (J, A) satisfies (3.2)–(3.5).

By subtracting (3.2)–(3.5) from (4.8)–(4.11), respectively, and using the pointwise convergence established above, we can show that $\|\underline{J}^k - J\|_\infty$, $\|\underline{A}^k - A\|_\infty \to 0$ as $k \to \infty$. Similarly, we can show that $\|\overline{J}^k - J\|_\infty$, $\|\overline{A}^k - A\|_\infty \to 0$ as $k \to \infty$. Hence, the proof is complete. □

From the aforementioned process, we also have the following comparison result.

Corollary 4.3. *Suppose that hypotheses* (A1)–(A6) *hold. Furthermore, suppose that* $(\underline{J}, \underline{A})$ *and* $(\overline{J}, \overline{A})$ *are a nonnegative lower solution and a nonnegative upper solution, respectively, of* (2.1). *Then the solution* (J, A) *of* (2.1) *satisfies*

$$\underline{J}(a,t) \leq J(a,t) \leq \overline{J}(a,t) \qquad a.e.\ in\ (0, a_{\max}) \times (0, T_0),$$
$$\underline{A}(x,t) \leq A(x,t) \leq \overline{A}(x,t) \qquad a.e.\ in\ (x_{\min}, x_{\max}) \times (0, T_0).$$

Furthermore, since $T_0 \equiv a_{\max}$, viewing $(J(a, T_0), A(x, T_0))$ as a new initial condition, we can easily extend the above-mentioned arguments to the interval $0 \leq t \leq T$ for any $T > 0$. Thus we have the following global existence result.

Theorem 4.4. *The solution* (J, A) *of problem* (2.1) *exists for* $0 \leq t < \infty$.

Remark 4.5. Corollary 4.3 plays a crucial role in understanding the long-time behavior of the model solution. In particular, if one constructs a suitable upper and/or lower solution, one is able to conclude whether the population goes extinct or survives. An important feature of this comparison approach is that it can handle time-dependent vital rates (see next section).

5. Population extinction and persistence.

In this section we study the long-time behavior of the solution of problem (2.1). In particular, we establish conditions on the parameters in (2.1) under which the population becomes extinct or survives in infinite time. Our asymptotic analysis is different from all other investigations of structured population models and relies on the comparison principle developed here and the construction of suitable pairs of lower and upper solutions. We first present the extinction result. To this end, we impose the following additional assumptions on the parameters:

(A7) There exists a positive constant δ such that

$$\nu(a,t) \geq \delta \quad \text{for } (a,t) \in [0,a_{\max}] \times [0,\infty), \quad g(x_{\min},t) \geq \delta \quad \text{for } t \in [0,\infty),$$

and

$$g_x(x,t) + \mu(x,t,0) \geq \delta \quad \text{for } (x,t) \in [x_{\min},x_{\max}] \times [0,\infty).$$

(A8) $\frac{1}{\inf_{t\in[0,\infty)} g(x_{\min},t)} \exp\big(-\int_0^{a_{\max}} \inf_{t\in[0,\infty)} \nu(q,t)dq\big) < 1$.

(A9) $\int_{x_{\min}}^{x_{\max}} \sup_{t\in[0,\infty)} \beta(x,t,0) \exp\big(-\int_{x_{\min}}^{x} \inf_{t\in[0,\infty)} \big[\frac{g_x(p,t)+\mu(p,t,0)}{g(p,t)}\big]dp\big)dx < 1$.

Remark 5.1. If the parameters are time-independent, then (A8)–(A9) reduce to

$$\frac{1}{g(x_{\min})} \exp\left(-\int_0^{a_{\max}} \nu(q)dq\right) < 1,$$

$$g(x_{\min}) \int_{x_{\min}}^{x_{\max}} \frac{\beta(x,0)}{g(x)} \exp\left(-\int_{x_{\min}}^{x} \frac{\mu(p,0)}{g(p)}dp\right) dx < 1.$$

Clearly, these conditions imply that the inherent net reproduction number of the model (2.1) is

$$R_0 = \exp\left(-\int_0^{a_{\max}} \nu(q)dq\right) \int_{x_{\min}}^{x_{\max}} \frac{\beta(x,0)}{g(x)} \exp\left(-\int_{x_{\min}}^{x} \frac{\mu(p,0)}{g(p)}dp\right) dx < 1. \tag{5.1}$$

THEOREM 5.2. *Suppose that hypotheses* (A7)–(A9) *hold. Then the solution* (J,A) *of problem* (2.1) *tends to zero uniformly in* $(a,x) \in [0,a_{\max}] \times [x_{\min},x_{\max}]$ *as* $t \to \infty$.

Proof. Let $\underline{J}(a,t) = 0$ and $\underline{A}(x,t) = 0$. We then introduce $\overline{J}(a,t) = M\mathcal{J}(a)e^{-\sigma t}$ and $\overline{A}(x,t) = M\mathcal{A}(x)e^{-\sigma t}$, where M,σ are positive constants to be determined, and $\mathcal{J},\mathcal{A}$ are continuously differentiable functions with $\mathcal{J}(0)=1, \mathcal{A}(x_{\min})=1$.

Clearly, $(\underline{J},\underline{A})$ is a lower solution. In order to ensure that $(\overline{J},\overline{A})$ is an upper solution, it suffices to require the following:

$$-\sigma\mathcal{J} + \mathcal{J}' + \nu(a,t)\mathcal{J} \geq 0 \quad \text{for } a \in (0,a_{\max}), \tag{5.2}$$

$$-\sigma\mathcal{A} + g_x(x,t)\mathcal{A} + g(x,t)\mathcal{A}' + \mu(x,t,0)\mathcal{A} \geq 0 \quad \text{for } x \in (x_{\min},x_{\max}), \tag{5.3}$$

$$1 \geq \int_{x_{\min}}^{x_{\max}} \beta(x,t,0)\mathcal{A}(x)dx \quad \text{and} \quad g(x_{\min},t) \geq \mathcal{J}(a_{\max}) \quad \text{for } t \in [0,\infty), \tag{5.4}$$

$$M\mathcal{J}(a) \geq J_0(a) \quad \text{for } a \in [0,a_{\max}] \quad \text{and} \quad M\mathcal{A}(x) \geq A_0(x) \quad \text{for } x \in [x_{\min},x_{\max}]. \tag{5.5}$$

In view of (A7), we choose $\sigma \leq \delta$ and let $\mathcal{J},\mathcal{A}$ satisfy

$$\mathcal{J}' + \left(\inf_{t\in[0,\infty)} \nu(a,t) - \sigma\right)\mathcal{J} = 0 \quad \text{for } a \in (0,a_{\max}),$$

$$\mathcal{A}' + \inf_{t\in[0,\infty)} \left(\frac{g_x(x,t) + \mu(x,t,0) - \sigma}{g(x,t)} \right) \mathcal{A} = 0 \quad \text{for } x \in (x_{\min}, x_{\max}).$$

The solutions of the above equations are given by

$$\mathcal{J}(a) = \exp\left(-\int_0^a \left(\inf_{t\in[0,\infty)} \nu(q,t) - \sigma \right) dq \right),$$

$$\mathcal{A}(x) = \exp\left(-\int_{x_{\min}}^x \inf_{t\in[0,\infty)} \left(\frac{g_x(p,t) + \mu(p,t,0) - \sigma}{g(p,t)} \right) dp \right),$$

and (5.2), (5.3) are valid. By virtue of (A8), we then choose σ so small that (5.4) holds. Finally, we choose M large enough such that

$$M\mathcal{J}(a_{\max}) \geq \sup_{[0,a_{\max}]} J_0(a) \quad \text{and} \quad M\mathcal{A}(x_{\max}) \geq \sup_{[x_{\min},x_{\max}]} A_0(x).$$

Hence, by Corollary 4.3, we have

$$0 \leq \lim_{t\to\infty} J(a,t) \leq \lim_{t\to\infty} M\mathcal{J}(a)e^{-\sigma t} \leq \lim_{t\to\infty} Me^{-\sigma t} = 0$$

and

$$0 \leq \lim_{t\to\infty} A(x,t) \leq \lim_{t\to\infty} M\mathcal{A}(x)e^{-\sigma t} \leq \lim_{t\to\infty} Me^{-\sigma t} = 0. \qquad \square$$

We now present the persistence result. For this purpose, we assume the following condition:

(A10) There exist positive values φ_0 and N such that

$$\sup_{[x_{\min},x_{\max}]\times[0,\infty)\times[\varphi_0,\infty)} [\beta(x,t,\varphi) - \mu(x,t,\varphi)] = -N < 0.$$

We show that the total population of adults is uniformly bounded.

Lemma 5.3. *There exists a positive value φ^* such that $0 \leq \varphi(t) \leq \varphi^*$ for $0 \leq t < \infty$.*

Proof. Let $\psi(t) = \int_0^{a_{\max}} J(a,t)da$. Integrating $(2.1)_1$ and $(2.1)_2$ with respect to a and x, respectively, and making use of $(2.1)_3$ and $(2.1)_4$, we obtain

$$(5.6) \qquad \begin{aligned}(\psi(t) + \varphi(t))' = &\int_{x_{\min}}^{x_{\max}} [\beta(x,t,\varphi(t)) - \mu(x,t,\varphi(t))]A(x,t)dx \\ &-\int_0^{a_{\max}} \nu(a,t)J(a,t)da.\end{aligned}$$

If there is a t_0 such that $\varphi(t_0) > \varphi_0$, then from (5.6) we have

$$\psi'(t_0) + \varphi'(t_0) \leq -N\varphi_0,$$

and it follows that either after a finite time t_1 $\varphi(t) \leq \varphi_0$ or for $t_0 \leq t < \infty$, $\varphi(t) \leq \psi(t_0) + \varphi(t_0)$. $\square$

We then impose the following additional assumptions on the parameters:

(A11) $g_x(x,t) + \mu(x,t,\varphi^*) > 0 \quad$ for $(x,t) \in [x_{\min}, x_{\max}] \times [0,\infty)$.

(A12) $\frac{1}{\sup_{t\in[0,\infty)} g(x_{\min},t)} \exp\big(-\int_0^{a_{\max}} \sup_{t\in[0,\infty)} \nu(q,t)dq\big) \geq 1$.

(A13) $\int_{x_{\min}}^{x_{\max}} \inf_{t\in[0,\infty)} \beta(x,t,\varphi^*) \exp\big(-\int_{x_{\min}}^{x} \sup_{t\in[0,\infty)} \big[\frac{g_x(p,t)+\mu(p,t,\varphi^*)}{g(p,t)}\big] dp\big) dx \geq 1$.

Remark 5.4. If the model parameters do not depend on t, then assumptions (A12)–(A13) imply that $R_0 > 1$ with R_0 given in (5.1).

THEOREM 5.5. *Suppose that hypotheses* (A10)–(A13) *hold and* $J_0(a) > 0$ *for* $a \in [0, a_{\max}]$, $A_0(x) > 0$ *for* $x \in [x_{\min}, x_{\max}]$. *Then the solution of problem* (2.1) *is uniformly persistent.*

Proof. We first introduce two new parameters $\tilde{\mu}(\cdot,\cdot,\varphi), \tilde{\beta}(\cdot,\cdot,\varphi) \in L^\infty((x_{\min}, x_{\max}) \times (0,T))$ that are nonnegative, continuously differentiable in φ, with $\tilde{\mu}_\varphi \geq 0$, $\tilde{\beta}_\varphi \leq 0$, and satisfy for every $(x,t) \in [x_{\min}, x_{\max}] \times [0,\infty)$

$$\tilde{\mu}(x,t,\varphi(t)) = \begin{cases} \mu(x,t,\varphi(t)) & \text{if } \varphi(t) \leq \varphi^* - \varepsilon, \\ \mu(x,t,\varphi^*) & \text{if } \varphi(t) \geq \varphi^* + \varepsilon \end{cases}$$

and

$$\tilde{\beta}(x,t,\varphi(t)) = \begin{cases} \beta(x,t,\varphi(t)) & \text{if } \varphi(t) \leq \varphi^* - \varepsilon, \\ \beta(x,t,\varphi^*) & \text{if } \varphi(t) \geq \varphi^* + \varepsilon, \end{cases}$$

where ε is a small positive constant. Clearly, the solution of (2.1) is the limit (as $\varepsilon \to 0$) to the solution of (2.1*), which is (2.1) with μ and β replaced with $\tilde{\mu}$ and $\tilde{\beta}$, respectively. From now on, we focus on problem (2.1*).

Set $\underline{J}(a,t) = m\mathcal{K}(a)$, $\underline{A}(x,t) = m\mathcal{B}(x)$ and $\overline{J}(a,t) = Me^{-\rho a}e^{\gamma t}$, $\overline{A}(x,t) = Me^{-\eta(x-x_{\min})}e^{\gamma t}$, where m, M, ρ, η, γ are positive constants to be determined and $\mathcal{K}, \mathcal{B}$ are continuously differentiable functions with $\mathcal{K}(0) = 1, \mathcal{B}(x_{\min}) = 1$.

If $\underline{J}, \underline{A}$ and $\overline{J}, \overline{A}$ are coupled lower and upper solutions of (2.1*), they satisfy the following:

$$\mathcal{K}' + \nu(a,t)\mathcal{K} \leq 0 \quad \text{for } a \in (0, a_{\max}), \tag{5.7}$$

$$g_x(x,t)\mathcal{B} + g(x,t)\mathcal{B}' + \tilde{\mu}(x,t,\infty)\mathcal{B} \leq 0 \quad \text{for } (x,t) \in (x_{\min}, x_{\max}) \times (0,\infty), \tag{5.8}$$

$$1 \leq \int_{x_{\min}}^{x_{\max}} \tilde{\beta}(x,t,\infty)\mathcal{B}(x)dx \quad \text{and} \quad g(x_{\min},t) \leq \mathcal{K}(a_{\max}) \quad \text{for } t \in [0,\infty), \tag{5.9}$$

$$m\mathcal{K}(a) \leq J_0(a) \ \text{ for } a \in [0, a_{\max}] \quad \text{and} \quad m\mathcal{B}(x) \leq A_0(x) \ \text{ for } x \in [x_{\min}, x_{\max}], \tag{5.10}$$

$$\gamma - \rho + \nu(a,t) \geq 0 \quad \text{for } (a,t) \in (0, a_{\max}) \times (0,\infty), \tag{5.11}$$

$$\gamma + g_x(x,t) - \eta g(x,t) + \tilde{\mu}(x,t,0) \geq 0 \quad \text{for } (x,t) \in (x_{\min}, x_{\max}) \times (0,\infty), \tag{5.12}$$

$$1 \geq \int_{x_{\min}}^{x_{\max}} \tilde{\beta}(x,t,0)e^{-\eta(x-x_{\min})}dx \quad \text{and} \quad g(x_{\min},t) \geq e^{-\rho a_{\max}} \quad \text{for } t \in [0,\infty), \tag{5.13}$$

$$Me^{-\rho a} \geq J_0(a) \ \text{ for } a \in [0, a_{\max}] \ \text{ and } \ Me^{-\eta(x-x_{\min})} \geq A_0(x) \ \text{ for } x \in [x_{\min}, x_{\max}]. \tag{5.14}$$

We first set up two equations for $\mathcal{K}$ and $\mathcal{B}$, respectively:

$$\mathcal{K}' + \sup_{t\in[0,\infty)} \nu(a,t)\mathcal{K} = 0 \quad \text{for } a \in (0, a_{\max})$$

and

$$\mathcal{B}' + \sup_{t\in[0,\infty)} \left(\frac{g_x(x,t) + \tilde{\mu}(x,t,\infty)}{g(x,t)} \right) \mathcal{B} = 0 \quad \text{for } x \in (x_{\min}, x_{\max}).$$

Solving these equations, we obtain

$$\mathcal{K}(a) = \exp\left(-\int_0^a \sup_{t\in[0,\infty)} \nu(q,t)dq \right),$$

$$\mathcal{B}(x) = \exp\left(-\int_{x_{\min}}^x \sup_{t\in[0,\infty)} \left(\frac{g_x(p,t) + \tilde{\mu}(p,t,\infty)}{g(p,t)} \right) dp \right),$$

and taking (A12), (A13) into account, we find that (5.7)–(5.9) hold. We then choose m small enough such that

$$m \le \inf_{[0,a_{\max}]} J_0(a) \quad \text{and} \quad m \le \inf_{[x_{\min},x_{\max}]} A_0(x).$$

We now turn to $\overline{J}, \overline{A}$. We first choose $\rho = \max\{0, -\ln(\inf_{t\in[0,\infty)} g(x_{\min},t))/a_{\max}\}$ and $\eta = \sup_{[x_{\min},x_{\max}]\times[0,\infty)} \tilde{\beta}(x,t,0)$ such that (5.13) is valid. We then choose γ so large that (5.11) and (5.12) hold. Finally, we choose M large enough such that

$$Me^{-\rho a_{\max}} \ge \sup_{[0,a_{\max}]} J_0(a) \quad \text{and} \quad Me^{-\eta(x_{\max}-x_{\min})} \ge \sup_{[x_{\min},x_{\max}]} A_0(x).$$

Hence, by Corollary 4.3, we have

$$m \exp\left(-\int_0^{a_{\max}} \sup_{t\in[0,\infty)} \nu(q,t)dq \right) \le J(a,t) \quad \text{for } (a,t) \in [0,a_{\max}] \times [0,\infty)$$

and

$$m \exp\left(-\int_{x_{\min}}^{x_{\max}} \sup_{t\in[0,\infty)} \left(\frac{g_x(p,t) + \tilde{\mu}(p,t,\infty)}{g(p,t)} \right) dp \right) \le A(x,t) \quad \text{for } (x,t) \in [x_{\min}, x_{\max}] \times [0,\infty).$$

Since $\lim_{\varepsilon\to 0} \tilde{\mu}(x,t,\varphi) = \mu(x,t,\varphi)$ for $\varphi \le \varphi*$, letting $\varepsilon \to 0$, the proof is now complete. □.

6. Concluding remarks. In this paper we have developed a new method for studying the long-time behavior of a nonautonomous age-size–structured model. The key idea in this method is the establishment of a comparison principle and the construction of a suitable pair of upper and lower solutions. Although comparison principles for local hyperbolic partial differential equations have been known for a long time [24], the first such principle for a nonlocal and linear size-structured model describing the evolution of a single population was developed by the authors in [3] and later extended to a nonlinear size-structured model in [4]. Here, we extend these comparison principles to a nonlinear nonautonomous juvenile-adult model, and, most importantly, we show for the first time that such a method can be used to provide conditions on the vital rates (which are related to the net reproduction number of the model (2.1) that result in population extinction or in uniform persistence of the population).

It is our hope that this method will present another approach for understanding the long-time behavior of such (complex) autonomous or nonautonomous models. Indeed, our current efforts are focused on the careful construction of upper and lower solutions that result in convergence of solutions to an equilibrium for the case of time-independent parameters (autonomous) or even for the case of asymptotically autonomous models [30].

Acknowledgment. The authors thank the referees for their constructive comments and helpful suggestions.

REFERENCES

[1] A. S. Ackleh, H. T. Banks, and K. Deng, *A difference approximation for a coupled system of nonlinear size-structured populations*, Nonlinear Anal., 50 (2002), pp. 727–748.

[2] A. S. Ackleh and P. DeLeenheer, *Discrete three-stage population model: Persistence and global stability results*, J. Biol. Dynamics, 2 (2008), pp. 415–427.

[3] A. S. Ackleh and K. Deng, *A monotone approximation for the nonautonomous size-structured population model*, Quart. Appl. Math., 57 (1999), pp. 261–267.

[4] A. S. Ackleh and K. Deng, *Existence-uniqueness of solutions for a nonlinear nonautonomous size-structured population model: An upper-lower solution approach*, Canadian Appl. Math. Quart., 8 (2000), pp. 1–15.

[5] A. S. Ackleh and K. Deng, *Survival of the fittest in a quasilinear size-structured population model*, Natural Resource Modelling, 17 (2004), pp. 213–228.

[6] A. S. Ackleh, K. Deng, and X. Wang, *Competitive exclusion and coexistence in a quasilinear size-structured population model*, Math. Biosci., 192 (2004), pp. 177–192.

[7] A. S. Ackleh, Y. Dib, and S. Jang, *A three-stage discrete-time population model: Seasonal versus continuous reproduction*, J. Biol. Dynamics, 1 (2007), pp. 305–319.

[8] A. S. Ackleh and K. Ito, *An implicit finite difference scheme for the nonlinear size-structured population model*, Numer. Funct. Anal. Optim., 18 (1997), pp. 865–884.

[9] A. S. Ackleh and S. Jang, *A discrete two-stage population model: Continuous versus seasonal reproduction*, J. Differ. Equations Appl., 13 (2007), pp. 261–274.

[10] R. Bellman, *Methods of Nonlinear Analysis*, Vol. II, Academic Press, New York, 1973.

[11] A. Calsina and J. Saldana, *A model of physiologically structured population dynamics with a nonlinear individual growth rate*, J. Math. Biol., 33 (1995), pp. 335–364.

[12] Á. Calsina and J. Saldaña, *Global dynamics and optimal life history of a structured population model*, SIAM J. Appl. Math., 59 (1999), pp. 1667–1685.

[13] J. M. Cushing, *The dynamics of hierarchical age-structured populations*, J. Math. Biol., 32 (1994), pp. 705–729.

[14] J. M. Cushing, *A juvenile-adult model with periodic vital rates*, J. Math. Biol., 53 (2006), pp. 520–539.

[15] J. M. Cushing and J. Li, *Juvenile versus adult competition*, J. Math. Biol., 29 (1991), pp. 457–473.

[16] J. P. Collins and A. Storfer, *Amphibian decline: Sorting the hypothesis*, Diversity Distributions, 9 (2003), pp. 89–98.

[17] O. Diekmann, P. Getto, and M. Gyllenberg, *Stability and bifurcation analysis of Volterra functional equations in the light of suns and stars*, SIAM J. Math. Anal., 39 (2007), pp. 1023–1069.

[18] O. Diekmann, M. Gyllenberg, H. Huang, M. Kirkilionis, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models.* II. *Nonlinear theory*, J. Math. Biol., 43 (2001), pp. 157–189.

[19] O. Diekmann, M. Gyllenberg, J. A. J. Metz, and H. R. Thieme, *On the formulation and analysis of general deterministic structured population models.* I. *Linear theory*, J. Math. Biol., 36 (1998), pp. 349–388.

[20] J. Farkas and T. Hagen, *Asymptotic behavior of size-structured population via juvenile-adult interaction*, Discrete Contin. Dynam. Systems Ser. B, 9 (2008), pp. 249–266.

[21] J. S. Garton and R. A. Brandon, *Reproductive ecology of the H. cinerea, Hyla cinerea, in southern Illinois (anura: hylidae)*, Herpetologica, 31 (1975), pp. 150–161.

[22] M. S. Gunzburger, *Reproductive ecology of the H. cinerea (Hyla cinerea) in northwestern Florida*, Amer. Midland Naturalist, 155 (2006), pp. 321–328.

[23] M. Iannelli, *Mathematical Theory of Age-Structured Population Dynamics*, Giardini Editori, Pisa, Italy, 1994.

[24] G. S. Ladde, V. Lakshmikantham, and A. S. Vatsala, *Monotone Iterative Techniques for Nonlinear Differential Equations*, Pitman, Boston, 1985.

[25] C. V. Pao, *Nonlinear Parabolic and Elliptic Equations*, Plenum Press, New York, 1992.

[26] L. Pham, S. Boudreaux, S. Karhbet, B. Price, A. S. Ackleh, J. Carter, and N. Pal, *Population estimates of Hyla cinerea (Schneider) (green tree frog) in an urban environment*, Southeastern Naturalist, 6 (2007), pp. 203–216.

[27] E. Shim, Z. Feng, C. Castillo-Chavez, and M. Martcheva, *An age-structured epidemic model of rotavirus with vaccination*, J. Math. Biol., 53 (2006), pp. 719–746.

[28] K. T. Smith, *Effects of nonindigenous tadpoles on native tadpoles in Florida: Evidence of competition*, Biol. Conservation, 123 (2005), pp. 433–441.

[29] A. Storfer, *Amphibian decline: Future directions*, Diversity Distributions, 9 (2003), pp. 151–163.

[30] H. Thieme, *Convergence results and a Poincare-Bendixson trichotomy for asymptotically autonomous differential equations*, J. Math. Biol., 30 (1992), pp. 755–763.

[31] J. R. Vonesh and O. D. L. Cruz, *Complex life cycles and density dependence: Assessing the contribution of egg mortality to amphibian declines*, Oecologia, 133 (2002), pp. 325–333.

[32] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Monogr. Textbooks Pure Appl. Math. 89, Marcel Dekker, New York, 1985.

[33] A. H. Wright and A. A. Wright, *Handbook of Frogs and Toads of the United States and Canada*, Comstock Publishing Company, Ithaca, NY, 1949.

SIAM J. Appl. Math.
Vol. 69, No. 6, pp. 1662–1681

DETECTING INCLUSIONS IN ELECTRICAL IMPEDANCE TOMOGRAPHY WITHOUT REFERENCE MEASUREMENTS*

BASTIAN HARRACH† AND JIN KEUN SEO‡

Abstract. We develop a new variant of the factorization method that can be used to detect inclusions in electrical impedance tomography from either absolute current-to-voltage measurements at a single, nonzero frequency or from frequency-difference measurements. This eliminates the need for numerically simulated reference measurements at an inclusion-free body and thus greatly improves the method's robustness against forward modeling errors, e.g., in the assumed body's shape.

Key words. inverse problems, electrical impedance tomography, complex conductivity, frequency-difference measurements, factorization method

AMS subject classifications. 35R30, 35Q60, 35J25, 35R05

DOI. 10.1137/08072142X

1. Introduction.

In electrical impedance tomography (EIT), we inject a time-harmonic current of I mA at a fixed angular frequency ω into an imaging subject using a pair of surface electrodes attached to its boundary. Then the induced time-harmonic electrical potential is dictated by the complex conductivity distribution σ^ω of the subject, the applied current, and the shape of the subject, where the real and imaginary part of the complex conductivity, $\Re(\sigma^\omega)$ and $\Im(\frac{\sigma^\omega}{\omega})$, are the real conductivity and the permittivity at the angular frequency ω, respectively. In EIT, we use measured boundary voltages generated by multiple injection currents to reconstruct an image of σ^ω inside the subject. It is well known that these boundary measurements are very insensitive and highly nonlinear to any local change of conductivity values away from the measuring points. Hence, the reconstructed image quality in terms of accuracy would be affected sensitively by unavoidable errors including the modeling errors and measurement noises.

Understanding the limited capabilities of static EIT imaging under realistic environments, numerous recent studies in EIT focus on the detection of conductivity anomalies instead of (e.g., cross-sectional) conductivity imaging; cf., e.g., [17, 21, 27, 28, 2, 9, 1, 20, 8, 15], the references therein, and the works connected with the factorization method cited further below.

Let us briefly explain the anomaly detection problem in EIT. Let the imaging object occupy a two- or three-dimensional region B with its smooth boundary ∂B, and let anomalies occupy a region Ω inside a background domain B of constant conductivity. We furthermore assume that the conductivity is isotropic. To distinguish the conductivity of the anomaly Ω and the surrounding homogeneous domain $B \setminus \Omega$, we denote the conductivity distribution at $\omega = 0$ by

$$\sigma(x) = \sigma_0 + \sigma_\Omega(x)\chi_\Omega(x),$$

*Received by the editors April 16, 2008; accepted for publication (in revised form) January 7, 2009; published electronically March 27, 2009.
http://www.siam.org/journals/siap/69-6/72142.html

†Birth name: Bastian Gebauer, Institut für Mathematik, Johannes Gutenberg-Universität, 55099 Mainz, Germany (harrach@math.uni-mainz.de). The work of this author was supported by the German Federal Ministry of Education and Research (BMBF) under grant 03HBPAM2, "Regularization techniques for electrical impedance tomography in medical and geological sciences".

‡Computational Science and Engineering, Yonsei University, Seoul, 120-749, Korea (seoj@yonsei.ac.kr). The work of this author was supported by WCU program R31-2008-000-10049-0.

where χ_Ω is the characteristic function of Ω and σ is a positive and bounded function in B. The inverse problem is to identify Ω from several pairs of Neumann-to-Dirichlet data

$$(g_j, \Lambda(g_j)) \in L^2_\diamond(\partial B) \times L^2_\diamond(\partial B), \quad j = 1, \dots, L,$$

where $L^2_\diamond(\partial B) = \{\phi \in L^2(\partial B) \ : \ \int_{\partial B} \phi \ \mathrm{d}x = 0\}$. Here, $\Lambda(g) = u|_{\partial B}$ and u is the $H^1(B)$-solution for the Neumann boundary value problem:

$$\begin{aligned} \nabla \cdot (\sigma \nabla u) &= 0 \qquad \text{in } B, \\ \sigma \frac{\partial u}{\partial \nu}|_{\partial B} &= g, \qquad \int_{\partial B} u \ \mathrm{d}x = 0, \end{aligned}$$

where ν is the unit outward normal vector to the boundary ∂B.

One of the most successful EIT-methods for locating multiple anomalies would be the factorization method introduced by Kirsch [24] for inverse scattering problems and generalized to EIT-problems by Brühl and Hanke in [5, 4]; see also the recent book of Kirsch and Grinberg [26], the work of Kirsch [25] on the complex conductivity case, and [6, 14, 18, 11, 16, 19, 30, 13, 29] for further extensions of the method in the context of EIT. The factorization method is based on a characterization using the range of the difference between the Neumann-to-Dirichlet (NtD) map in the presence of anomalies and that in the absence of anomalies: $z \in \Omega$ if and only if $\Phi_z|_{\partial B}$ is in the range of the operator $|\Lambda - \Lambda_0|^{1/2}$, where Λ_0 is the NtD map corresponding to the reference homogeneous conductivity $\sigma(x) = \sigma_0$ and $\Phi_z(x)$ is the solution of

$$\Delta_x \Phi_z(x) = d \cdot \nabla_x \delta_z(x) \quad \text{ in } B, \quad \frac{\partial}{\partial \nu}\Phi_z|_{\partial\Omega} = 0, \quad \text{ and } \quad \int_{\partial B} \Phi_z(x) \ \mathrm{d}x = 0,$$

where d is any unit vector and δ_z is the Dirac delta function at z.

For the practical application of the factorization method for static EIT systems, the requirement of the reference NtD data Λ_0 is a drawback. While, in practice, a rough approximation of the NtD map Λ can be obtained from the current-to-voltage data, the corresponding current-to-voltage data for the reference NtD map Λ_0 in the absence of anomalies is usually not available.

Hence, one uses numerically simulated data corresponding to Λ_0 by solving the forward problem $\nabla \cdot (\sigma_0 \nabla u) = 0$ in B with mimicked Neumann data representing the injection current in the EIT system. Noting that the simulated Dirichlet data is mainly depending on the geometry of ∂B and the Neumann data, instead of the conductivity σ_0 (which acts merely as a scaling factor), the requirement of the reference NtD data Λ_0 makes the factorization method very sensitive to forward modeling errors including the boundary geometry error and electrodes position uncertainty (related to the mimicked Neumann data), since its image reconstruction problem is ill-posed. Hence, it is desirable to eliminate the requirement of the reference NtD data Λ_0.

In this work, we adopt the frequency-difference EIT system [31, 32] to obtain a subsidiary NtD data Λ_ω at a fixed angular frequency ω taken from the range of 1kHz $\leq \frac{\omega}{2\pi} \leq$ 500kHz. $\Lambda_\omega(g)$ is the Dirichlet data of the complex potential u_ω which satisfies

$$\nabla \cdot (\sigma^\omega \nabla u_\omega) = 0 \quad \text{ in } B, \quad \sigma^\omega \frac{\partial}{\partial \nu} u_\omega|_{\partial B} = g, \quad \text{ and } \quad \int_{\partial B} u_\omega \ \mathrm{d}x = 0.$$

Our aim is to substitute Λ_ω for Λ_0 in the conventional factorization method and use an interrelation between Λ and Λ_ω to locate the anomalies Ω. However, due to

$\Im\{\sigma^\omega\} \neq 0$, the operator Λ_ω is not self-adjoint and $\Lambda - \Lambda_\omega$ is neither semipositive nor seminegative.

In this work, we show that, for an arbitrary fixed nonzero frequency ω, both, the imaginary part of $\sigma_0^\omega \Lambda_\omega$ and the real part of the normalized difference $\sigma_0 \Lambda - \sigma_0^\omega \Lambda_\omega$ (or actually any other normalized difference of measurements taken at two different frequencies), provide a constructive way of locating Ω, where σ_0^ω is the background complex conductivity at the angular frequency ω. To our knowledge this is the first characterization result that works without reference measurements. We numerically demonstrate that the proposed new variant of the factorization method locates successfully the region Ω with a reasonable accuracy in the presence of boundary geometry errors and measurement noise. We also describe a heuristic approach to estimate an unknown background conductivity from the measured data and numerically test it on a homogeneous, as well as on a slightly inhomogeneous, background.

In section 2 we formulate and prove our main results. In section 3 we test our method numerically, compare its sensitivity to body shape errors with the conventional factorization method, and describe a heuristic approach to estimate an unknown background conductivity. Section 4 contains some concluding remarks.

2. Characterization of an inclusion without reference data.

Let $B \subset \mathbb{R}^n$, $n \geq 2$, be a smoothly bounded domain describing the investigated body. Let $\omega > 0$ be an arbitrary fixed frequency and denote by σ^ω the body's complex conductivity at some fixed nonzero frequency $\omega > 0$. We assume that $\Re(\sigma^\omega) \in L^\infty_+(B;\mathbb{R})$ and $\Im(\sigma^\omega) \in L^\infty(B;\mathbb{R})$, where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary part, the subscript "+" indicates functions with positive (essential) infima, and throughout this work all function spaces consist of complex valued functions if not stated otherwise.

A time-harmonic current with (complex) amplitude $g \in L^2_\diamond(\partial B)$ and frequency ω that is applied to the body's surface gives rise to an electric potential $u_\omega \in H^1(B)$ that satisfies

$$\nabla \cdot (\sigma^\omega \nabla u_\omega) = 0 \quad \text{in } B \quad \text{and} \quad \sigma^\omega \partial_\nu u_\omega|_{\partial B} = g, \tag{2.1}$$

where $L^2_\diamond(\partial B)$ is the subspace of $L^2(\partial B)$-functions with vanishing integral mean, ν is the outer normal on ∂B.

It is a well-known consequence of the Lax–Milgram theorem (cf., e.g., [7, Chapter VI, §3, Theorem 7]) that there exists a solution of (2.1) and that this solution is uniquely determined up to addition of a constant function. We denote the quotient space of $H^1(B)$ modulo constant functions by $H^1_\diamond(B)$. The trace operator $v \mapsto v|_{\partial B}$ canonically extends to $H^1_\diamond(B) \to L^2(\partial B)/\mathbb{C}$, where we identify the latter space with $L^2_\diamond(\partial B)$ by appropriately fixing the ground level.

The inverse problem of frequency-dependent EIT is the problem of determining (properties of) σ^ω from measuring one or several pairs of Neumann and Dirichlet boundary values $(u_\omega|_{\partial B}, \sigma^\omega \partial_\nu u_\omega|_{\partial B})$. Mathematically, the knowledge of all such pairs is equivalent to knowing the Neumann-to-Dirichlet operator

$$\Lambda_\omega : L^2_\diamond(\partial B) \to L^2_\diamond(\partial B), \quad g \mapsto u_\omega|_{\partial B},$$

where u_ω solves (2.1). It is easily checked that Λ_ω is linear and compact.

2.1. The main results.

In this work, we assume that the conductivity of the body is constant outside one or several inclusions, i.e.,

$$\sigma^\omega(x) = \sigma_0^\omega + \sigma_\Omega^\omega(x)\chi_\Omega(x),$$

where Ω is some open (possibly disconnected) set with smooth boundary and connected complement, $\overline{\Omega} \subset B$, and $\sigma_0^\omega \in \mathbb{C}$, $\sigma_\Omega^\omega \in L^\infty(\Omega)$ are such that $\Re(\sigma^\omega) \in L_+^\infty(B;\mathbb{R})$.

We will show that the inclusion Ω can be determined from the Neumann-to-Dirichlet operator Λ_ω using the same singular dipole potentials that were introduced for the factorization method by Brühl and Hanke in [5, 4]. For an arbitrary direction $d \in \mathbb{R}^n$, $|d| = 1$, and every point $z \in B$, let Φ_z be the solution of

$$\Delta_x \Phi_z(x) = d \cdot \nabla_x \delta_z(x) \quad \text{in } B$$

with homogeneous Neumann boundary values $\partial_\nu \Phi_z(x)|_{\partial B} = 0$ and vanishing integral mean on ∂B.

THEOREM 2.1. *Assume that either*

$$\Im\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) \in L_+^\infty(\Omega;\mathbb{R}) \quad \textit{or} \quad -\Im\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) \in L_+^\infty(\Omega;\mathbb{R}). \tag{2.2}$$

Then

$$z \in \Omega \quad \textit{if and only if} \quad \Phi_z|_{\partial B} \in \mathcal{R}\left(\left|\Im\left(\sigma_0^\omega \Lambda_\omega\right)\right|^{1/2}\right),$$

where the imaginary part of a bounded linear operator $A \in \mathcal{L}(H)$ on a complex Hilbert space H is defined by $\Im(A) := \frac{1}{2\mathrm{i}}(A - A^*)$, and $\mathcal{R}(A)$ denotes the range of A.

We also show a complementary result for the real part of frequency-difference data. Let $0 \leq \tau \neq \omega$ be another fixed frequency (possibly being zero) for which the body's complex conductivity is σ^τ, with $\Re(\sigma^\tau) \in L_+^\infty(B;\mathbb{R})$. We assume that σ^τ is also constant outside the same inclusion Ω, i.e., $\sigma^\tau = \sigma_0^\tau + \sigma_\Omega^\tau(x)\chi_\Omega(x)$ with $\sigma_0^\tau \in \mathbb{C}$. Measurements at the frequency τ are described by the NtD operator Λ_τ which is defined analogously to Λ_ω.

THEOREM 2.2. *If either*

$$\Re\left(\tfrac{\sigma_\Omega^\tau}{\sigma_0^\tau}\right) - \Re\left(\tfrac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) - \frac{\Im\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right)^2}{\Re\left(\frac{\sigma^\omega}{\sigma_0^\omega}\right)} \in L_+^\infty(\Omega;\mathbb{R}), \tag{2.3}$$

or

$$\Re\left(\tfrac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) - \Re\left(\tfrac{\sigma_\Omega^\tau}{\sigma_0^\tau}\right) - \frac{\Im\left(\frac{\sigma_\Omega^\tau}{\sigma_0^\tau}\right)^2}{\Re\left(\frac{\sigma^\tau}{\sigma_0^\tau}\right)} \in L_+^\infty(\Omega;\mathbb{R}), \tag{2.4}$$

then

$$z \in \Omega \quad \textit{if and only if} \quad \Phi_z|_{\partial B} \in \mathcal{R}\left(\left|\Re\left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau\right)\right|^{1/2}\right),$$

where the real part of a bounded linear operator $A \in \mathcal{L}(H)$ on a complex Hilbert space H is defined by $\Re(A) := \frac{1}{2}(A + A^*)$.

Before we prove these two theorems in the next subsection, let us comment on their relevance. In contrast to the conventional factorization method where the measured NtD data is compared to reference data that is usually not available by experiment (with the disadvantages described in the introduction), both theorems use only experimentally available NtD measurements. Our theorems require NtD data either at just one (nonzero) frequency, which is then compared to its own adjoint (Theorem 2.1) or NtD data measured at two different frequencies (possibly one being zero), which are then compared to each other (Theorem 2.2). In particular, this means that we can replace unavailable reference measurements in the conventional factorization methods by experimentally available measurements at an arbitrary nonzero frequency, which strongly reduces the methods sensitivity to boundary geometry errors as we will demonstrate numerically in section 3. In practice, one may have access to measurements at more than two frequencies. This redundancy can surely be used to further increase the performance or robustness of our method, but we have not studied this question in detail.

The assumptions (2.2)–(2.4) are arguably not very intuitive. For the practically relevant model case that the real conductivity $\kappa = \Re(\sigma^\omega)$ and the permittivity $\epsilon = \Im(\frac{\sigma^\omega}{\omega})$ are not frequency-dependent, they can be restated as follows.

Remark 2.3. Let

$$\sigma^\tau(x) = \kappa(x) + \mathrm{i}\tau\epsilon(\mathrm{x}), \qquad \kappa(x) = \kappa_0 + \kappa_\Omega(x)\chi_\Omega(x),$$

$$\sigma^\omega(x) = \kappa(x) + \mathrm{i}\omega\epsilon(\mathrm{x}), \qquad \epsilon(x) = \epsilon_0 + \epsilon_\Omega(x)\chi_\Omega(x),$$

with $\kappa_0, \epsilon_0 \in \mathbb{R}$, $\kappa_\Omega, \epsilon_\Omega \in L^\infty(\Omega;\mathbb{R})$. Then

$$\Im\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) = \frac{\omega}{\kappa_0^2 + \omega^2\epsilon_0^2}(\kappa_0\epsilon_\Omega - \kappa_\Omega\epsilon_0),$$

$$\Re\left(\frac{\sigma_\Omega^\tau}{\sigma_0^\tau}\right) - \Re\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right) - \frac{\Im\left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega}\right)^2}{\Re\left(\frac{\sigma^\omega}{\sigma_0^\omega}\right)} = \frac{(\kappa_0\epsilon_\Omega - \kappa_\Omega\epsilon_0)(\tau^2\kappa\epsilon_0 - \omega^2\kappa_0\epsilon)}{(\kappa_0^2 + \tau^2\epsilon_0^2)(\kappa_0\kappa + \omega^2\epsilon_0\epsilon)},$$

and the same identities hold with ω and τ interchanged. Hence, (2.2) is equivalent to

$$\kappa_0\epsilon_\Omega - \kappa_\Omega\epsilon_0 \in L_+^\infty(\Omega;\mathbb{R}) \quad \text{or} \quad \kappa_\Omega\epsilon_0 - \kappa_0\epsilon_\Omega \in L_+^\infty(\Omega;\mathbb{R}). \tag{2.5}$$

If ω is sufficiently larger than τ, then (2.5) is also equivalent to the disjunction of (2.3) and (2.4). For $\tau = 0$, every $\omega > 0$ is sufficiently large.

In particular, our method can identify inclusions where only the real conductivity or only the permittivity differs from the background, and this deviance is of the same sign in all inclusions. However, there exist combinations where a jump κ_Ω in the real conductivity and a jump ϵ_Ω in the permittivity cancel each other out, in the sense that

$$\kappa_0\epsilon_\Omega - \kappa_\Omega\epsilon_0 = 0.$$

Inclusions with this property cannot be detected by our method. In fact, they are completely invisible to weighted frequency-difference measurements, as one easily checks that in this case

$$\Im\left(\sigma_0^\omega\Lambda_\omega\right) = 0 \quad \text{and} \quad \Re\left(\sigma_0^\omega\Lambda_\omega - \sigma_0^\tau\Lambda_\tau\right) = 0.$$

There is also another drawback in our method compared to the original factorization method. The original method also works in the case of an inhomogeneous (but known) background medium, which only requires the replacement of the singular dipole potentials Φ_z by the corresponding dipole potentials in the inhomogeneous background. For our method this is not sufficient, since we also need the constant background conductivities to form the weighted differences of the NtD data. Hence, up to now, our method can only be applied to a homogeneous (constant) background medium. We believe, however, that our results cover the most relevant case in practice, where no accurate information about the background is available (except of being "almost constant"), and we describe in subsection 3.3 a heuristic approach how even an unknown background value can be estimated from the measured data. Our numerical results indicate that this approach still yields reasonable results if the unknown background is slightly inhomogeneous.

2.2. Proof of the main results. We start with some notations. The inner product on a complex Hilbert space H is denoted by $(\cdot,\cdot)$ and the pairing on H and its antidual H' is denoted by $\langle\cdot,\cdot\rangle_{H'\times H}$. Both are ordered in the way that they are linear in the first and antilinear in their second argument. For an operator $A\in\mathcal{L}(H_1,H_2)$ acting between Hilbert spaces H_1 and H_2, we rigorously distinguish between the dual operator $A'\in\mathcal{L}(H_2',H_1')$ and the adjoint operator $A^*\in\mathcal{L}(H_2,H_1)$.

Analogously to the definition of $H^1_\diamond(B)$, we denote by $H^1_\diamond(B\setminus\overline{\Omega})$, $H^1_\diamond(\Omega)$, and $H^{1/2}_\diamond(\partial\Omega)$ the quotient spaces of $H^1(B\setminus\overline{\Omega})$, $H^1(\Omega)$, and of the the trace space $H^{1/2}(\partial\Omega)$ modulo locally constant functions. (Thus, for disconnected Ω a multidimensional space is factored out.) The antidual of $H^{1/2}_\diamond(\partial\Omega)$ is denoted by $H^{-1/2}_\diamond(\partial\Omega)$, and we identify $L^2_\diamond(\partial B)$ with its antidual.

We will prove both results using a factorization of the difference of $\sigma_0^\omega\Lambda_\omega$ and the reference NtD operator Λ_0 corresponding to a constant conductivity equal to one, i.e.,

$$\Lambda_0:\ L^2_\diamond(\partial B)\to L^2_\diamond(\partial B),\quad g\mapsto u_0|_{\partial B},$$

where u solves

$$\Delta u_0=0\quad\text{ in } B\quad\text{ and }\quad \partial_\nu u_0|_{\partial B}=g.$$

It is well known and easily checked that Λ_0 is linear, compact, and self-adjoint. Let us stress again that we only require measurements at one arbitrary fixed frequency $\omega>0$ (for Theorem 2.1), or at two arbitrary, but different, fixed frequencies $\omega>0$, $\tau\geq 0$ (for Theorem 2.2). Λ_0 merely serves as an auxiliary operator that will cancel out later in our proof. Unlike other applications of the factorization method, Λ_0 does not have to correspond to real measurements.

For the factorization we also introduce the operator

$$L:\ H^{-1/2}_\diamond(\partial\Omega)\to L^2_\diamond(\partial B),\quad \psi\mapsto w|_{\partial B},$$

where $w\in H^1_\diamond(B\setminus\overline{\Omega})$ solves

$$\Delta w=0,\qquad \partial_\nu w|_{\partial B}=0,\qquad \partial_\nu w^+|_{\partial\Omega}=-\psi,$$

with ν being the normal on $\partial\Omega$ oriented into $B\setminus\overline{\Omega}$ and we denote by the superscripts "+", resp., "−" that the trace is taken from $B\setminus\overline{\Omega}$ and Ω, respectively. We also define

$$F_0:\ H^{1/2}_\diamond(\partial\Omega)\to H^{-1/2}_\diamond(\partial\Omega),\qquad F_0\phi=\partial_\nu v_0^+|_{\partial\Omega},$$

$$F_\omega:\ H^{1/2}_\diamond(\partial\Omega)\to H^{-1/2}_\diamond(\partial\Omega),\qquad F_\omega\phi=\partial_\nu v_\omega^+|_{\partial\Omega},$$

where $v_0, v_\omega \in H^1_\diamond(B \setminus \partial\Omega)$ solve

$$\begin{aligned} \Delta v_0 &= 0 \quad \text{in } B \setminus \partial\Omega, & \nabla \cdot \frac{\sigma^\omega}{\sigma_0^\omega} \nabla v_\omega &= 0 \quad \text{in } B \setminus \partial\Omega, \\ \partial_\nu v_0|_{\partial B} &= 0, & \partial_\nu v_\omega|_{\partial B} &= 0, \\ [v_0]_{\partial\Omega} &= \phi, & [v_\omega]_{\partial\Omega} &= \phi, \\ [\partial_\nu v_0]_{\partial\Omega} &= 0, & \left[\frac{\sigma^\omega}{\sigma_0^\omega} \partial_\nu v_\omega\right]_{\partial\Omega} &= 0, \end{aligned}$$

with $[\cdot]$ denoting the difference of the trace taken from $B \setminus \overline{\Omega}$ minus the trace taken from Ω. (Note that one easily checks that F_0 and F_ω are indeed well defined on these spaces.)

For the case of frequency-difference data, we also define F_τ analogously to F_ω. For the sake of readability we formulate the next two lemmas only for the frequency ω, though they just as well hold for τ.

LEMMA 2.4. *The difference of the Neumann-to-Dirichlet operators can be factorized into*

$$\Lambda_0 - \sigma_0^\omega \Lambda_\omega = L(F_0 - F_\omega)L'.$$

Proof. We proceed similarly to [10]. For given $g \in H_\diamond^{-\frac{1}{2}}(\partial B)$ let $\tilde{w} \in H^1_\diamond(B \setminus \overline{\Omega})$ solve

$$\Delta \tilde{w} = 0 \quad \text{in } B \setminus \overline{\Omega} \quad \text{and} \quad \partial_\nu \tilde{w} = \begin{cases} 0 & \text{on } \partial\Omega, \\ g & \text{on } \partial B. \end{cases}$$

Let $\psi \in H_\diamond^{-\frac{1}{2}}(\partial\Omega)$ and $w \in H^1_\diamond(B \setminus \overline{\Omega})$ be the function from the definition of $L\psi$. Then

$$\begin{aligned} \langle \psi, L'g \rangle = \overline{\langle g, L\psi \rangle} = \overline{\langle \partial_\nu \tilde{w}|_{\partial B}, w|_{\partial B} \rangle} = \int_{B \setminus \overline{\Omega}} \overline{\nabla \tilde{w}} \cdot \nabla w \, \mathrm{d}x = \left\langle -\partial_\nu w^+|_{\partial\Omega}, \tilde{w}^+|_{\partial\Omega} \right\rangle \\ = \left\langle \psi, \tilde{w}^+|_{\partial\Omega} \right\rangle, \end{aligned}$$

and thus $L'g = \tilde{w}^+|_{\partial\Omega}$.

Now let $v_0, v_\omega \in H^1_\diamond(B \setminus \partial\Omega)$ be the solutions from the definition of $F_0 \tilde{w}^+|_{\partial\Omega}$, resp., $F_\omega \tilde{w}^+|_{\partial\Omega}$. We define $u_0, u_\omega \in H^1_\diamond(B \setminus \partial\Omega)$ by setting $u_0 = -v_0$, resp., $u_\omega = -v_\omega$ on Ω and $u_0 = \tilde{w} - v_0$, resp., $u_\omega = \tilde{w} - v_\omega$ on $B \setminus \overline{\Omega}$. Then $u_0, \frac{1}{\sigma_0^\omega} u_\omega \in H^1_\diamond(B)$ and they solve the equations in the definitions of $\Lambda_0 g$ and $\Lambda_\omega g$. Thus,

$$(\Lambda_0 - \sigma_0^\omega \Lambda_\omega) g = (u_0 - u_\omega)|_{\partial B} = -(v_0 - v_\omega)|_{\partial B}.$$

Since $\Delta(v_0 - v_\omega) = 0$ in $B \setminus \overline{\Omega}$ and $\partial_\nu (v_0 - v_\omega)|_{\partial B} = 0$, we also have

$$L(\partial_\nu (v_0^+ - v_\omega^+)|_{\partial\Omega}) = -(v_0 - v_\omega)|_{\partial B},$$

and thus

$$(\Lambda_0 - \sigma_0^\omega \Lambda_\omega) g = L(\partial_\nu (v_0^+ - v_\omega^+)|_{\partial\Omega}) = L(F_0 - F_\omega) \tilde{w}^+|_{\partial\Omega} = L(F_0 - F_\omega) L' g. \qquad \square$$

Lemma 2.5. *For given $\phi \in H_\diamond^{1/2}(\partial\Omega)$ let $v_0, v_\omega \in H_\diamond^1(B \setminus \partial\Omega)$ be the solutions in the definition of F_0, F_ω and let $v_\phi \in H^1(B \setminus \partial\Omega)$ be such that $v_\phi^+|_{\partial\Omega} = \phi$ and $v_\phi|_\Omega = 0$. Set $\tilde{v}_0 := v_0 - v_\phi$ and $\tilde{v}_\omega := v_\omega - v_\phi$. Then*

$$\langle (F_0 - F_\omega)\phi, \phi\rangle = \int_B |\nabla \tilde{v}_0|^2 \, \mathrm{d}x - \int_B \frac{\overline{\sigma^\omega}}{\overline{\sigma_0^\omega}} |\nabla \tilde{v}_\omega|^2 \, \mathrm{d}x.$$

Furthermore, there exists a constant $c_\omega > 0$ such that

$$\int_\Omega |\nabla \tilde{v}_\omega|^2 \, \mathrm{d}x = \int_\Omega |\nabla v_\omega|^2 \, \mathrm{d}x \geq c_\omega \|\phi\|^2 \quad \textit{for all } \phi \in H_\diamond^{1/2}(\partial\Omega).$$

Proof. One easily checks that the functions $\tilde{v}_0, \tilde{v}_\omega \in H_\diamond^1(B)$ solve

$$\int_B \nabla \tilde{v}_0 \cdot \overline{\nabla w} \, \mathrm{d}x = -\int_{B\setminus\Omega} \nabla v_\phi \cdot \overline{\nabla w} \, \mathrm{d}x, \tag{2.6}$$

$$\int_B \frac{\sigma^\omega}{\sigma_0^\omega} \nabla \tilde{v}_\omega \cdot \overline{\nabla w} \, \mathrm{d}x = -\int_{B\setminus\Omega} \nabla v_\phi \cdot \overline{\nabla w} \, \mathrm{d}x \tag{2.7}$$

for all $w \in H_\diamond^1(B)$. Thus, we obtain

$$\begin{aligned}
\langle (F_0 - F_\omega)\phi, \phi\rangle &= \left\langle \partial_\nu v_0^+|_{\partial\Omega}, \phi \right\rangle - \left\langle \partial_\nu v_\omega^+|_{\partial\Omega}, \phi \right\rangle \\
&= \int_{B\setminus\overline{\Omega}} \nabla v_\omega \cdot \overline{\nabla v_\phi} \, \mathrm{d}x - \int_{B\setminus\overline{\Omega}} \nabla v_0 \cdot \overline{\nabla v_\phi} \, \mathrm{d}x \\
&= \int_{B\setminus\overline{\Omega}} \nabla \tilde{v}_\omega \cdot \overline{\nabla v_\phi} \, \mathrm{d}x - \int_{B\setminus\overline{\Omega}} \nabla \tilde{v}_0 \cdot \overline{\nabla v_\phi} \, \mathrm{d}x \\
&= \int_B |\nabla \tilde{v}_0|^2 \, \mathrm{d}x - \int_B \overline{\left(\frac{\sigma^\omega}{\sigma_0^\omega}\right)} |\nabla \tilde{v}_\omega|^2 \, \mathrm{d}x.
\end{aligned}$$

To prove the second assertion we first note that the Neumann boundary values

$$F_\omega \phi = \partial_\nu v_\omega^+|_{\partial\Omega} = \frac{\sigma^\omega}{\sigma_0^\omega} \partial_\nu v_\omega^-|_{\partial\Omega}$$

depend continuously on $v_\omega|_\Omega = \tilde{v}_\omega|_\Omega \in H_\diamond^1(\Omega)$, so that there exists $c'_\omega > 0$ such that

$$\int_\Omega |\nabla \tilde{v}_\omega|^2 \, \mathrm{d}x = \int_\Omega |\nabla v_\omega|^2 \, \mathrm{d}x \geq c'_\omega \|F_\omega \phi\|^2 \quad \text{for all } \phi \in H_\diamond^{1/2}(\partial\Omega).$$

Thus, it only remains to show that F_ω is bijective. Its injectivity is obvious and its surjectivity is shown as in the proof of [10, Lemma 3.3] by checking that a right inverse of F_ω is given by $-\lambda_{B\setminus\overline{\Omega}}^\omega - \lambda_\Omega^\omega$, where λ_Ω^ω and $\lambda_{B\setminus\overline{\Omega}}^\omega$ are the NtD operators on the inclusion Ω, resp., on its complement $B \setminus \overline{\Omega}$. □

Lemma 2.6. *The assumptions of Theorem* 2.1 *imply that there exist $c, C > 0$ such that*

$$c\|L'g\|^2 \leq (|\Im\,(\sigma_0^\omega \Lambda_\omega)|\, g, g) \leq C\|L'g\|^2. \tag{2.8}$$

For the case of frequency-difference data, the assumptions of Theorem 2.2 *imply that there exist* $c', C' > 0$ *such that*

$$c' \left\| L'g \right\|^2 \leq \left(\left| \Re \left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau \right) \right| g, g \right) \leq C' \left\| L'g \right\|^2. \tag{2.9}$$

Proof. For every $g \in L^2_\diamond(\partial B)$, $\phi := L'g$, let $\tilde{v}_\omega, v_\phi$ be the functions defined in Lemma 2.5. Noting that $\Lambda_0 = \Lambda_0'$ and $F_0 = F_0'$, we obtain from Lemmas 2.4 and 2.5 that

$$\begin{aligned} \left(\Im \left(\sigma_0^\omega \Lambda_\omega \right) g, g \right) &= \left\langle \Im \left(F_\omega \right) L'g, L'g \right\rangle \\ &= \frac{1}{2\mathrm{i}} \left(\langle (F_\omega - F_0)\phi, \phi \rangle - \overline{\langle (F_\omega - F_0)\phi, \phi \rangle} \right) \\ &= - \int_B \Im \left(\frac{\sigma^\omega}{\sigma_0^\omega} \right) \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x = - \int_\Omega \Im \left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega} \right) \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x. \end{aligned}$$

Hence, $\left| \Im \left(\sigma_0^\omega \Lambda_\omega \right) \right|$ is either $\Im \left(\sigma_0^\omega \Lambda_\omega \right)$ or $-\Im \left(\sigma_0^\omega \Lambda_\omega \right)$, the lower bound in assertion (2.8) follows from Lemma 2.5, and the upper bound follows from the the continuity of F_ω.

To prove the second assertion for the case of frequency-difference data, let $\tilde{v}_\tau$ be defined analogously to $\tilde{v}_\omega$. We now proceed similar to Ide et al. [20, Lemma 2.6]; cf. also the similar arguments in Kang, Seo, and Sheen [23] and Kirsch [25]. Using (2.7) and its analog for the frequency τ, we derive

$$\begin{aligned} 0 &\leq \int_B \Re \left(\tfrac{\sigma^\tau}{\sigma_0^\tau} \right) \left| \nabla \tilde{v}_\tau - \frac{\frac{\sigma^\omega}{\sigma_0^\omega}}{\Re \left(\frac{\sigma^\tau}{\sigma_0^\tau} \right)} \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x \\ &= \int_B \Re \left(\tfrac{\sigma^\tau}{\sigma_0^\tau} \right) \left| \nabla \tilde{v}_\tau \right|^2 \, \mathrm{d}x - 2\Re \left(\int_B \frac{\sigma^\omega}{\sigma_0^\omega} \nabla \tilde{v}_\omega \cdot \overline{\nabla \tilde{v}_\tau} \, \mathrm{d}x \right) + \int_B \frac{\left| \frac{\sigma^\omega}{\sigma_0^\omega} \right|^2}{\Re \left(\frac{\sigma^\tau}{\sigma_0^\tau} \right)} \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x \\ &= - \int_B \Re \left(\tfrac{\sigma^\tau}{\sigma_0^\tau} \right) \left| \nabla \tilde{v}_\tau \right|^2 \, \mathrm{d}x + \int_B \frac{\left| \frac{\sigma^\omega}{\sigma_0^\omega} \right|^2}{\Re \left(\frac{\sigma^\tau}{\sigma_0^\tau} \right)} \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x. \end{aligned}$$

Thus, it follows from Lemmas 2.4 and 2.5 that

$$\begin{aligned} &\left(\Re \left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau \right) g, g \right) \\ &= \int_B \Re \left(\tfrac{\sigma^\omega}{\sigma_0^\omega} \right) \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x - \int_B \Re \left(\tfrac{\sigma^\tau}{\sigma_0^\tau} \right) \left| \nabla \tilde{v}_\tau \right|^2 \, \mathrm{d}x \\ &\geq \int_B \left(\Re \left(\tfrac{\sigma^\omega}{\sigma_0^\omega} \right) - \frac{\left| \frac{\sigma^\omega}{\sigma_0^\omega} \right|^2}{\Re \left(\frac{\sigma^\tau}{\sigma_0^\tau} \right)} \right) \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x \\ &= \int_\Omega \frac{\Re \left(\frac{\sigma^\omega}{\sigma_0^\omega} \right)}{\Re \left(\frac{\sigma^\tau}{\sigma_0^\tau} \right)} \left(\Re \left(\tfrac{\sigma_\Omega^\tau}{\sigma_0^\tau} \right) - \Re \left(\tfrac{\sigma_\Omega^\omega}{\sigma_0^\omega} \right) - \frac{\Im \left(\frac{\sigma_\Omega^\omega}{\sigma_0^\omega} \right)^2}{\Re \left(\frac{\sigma^\omega}{\sigma_0^\omega} \right)} \right) \left| \nabla \tilde{v}_\omega \right|^2 \, \mathrm{d}x. \end{aligned}$$

Using $\Re\left(\frac{\sigma^\omega}{\sigma_0^\omega}\right), \Re\left(\frac{\sigma^\tau}{\sigma_0^\tau}\right) \in L_+^\infty(B)$, and assumption (2.3), we obtain a $c'' > 0$ with

$$(\Re\left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau\right) g, g) \geq c'' \int_\Omega \left|\nabla \tilde{v}_\omega\right|^2 \, \mathrm{d}x.$$

An analog equation follows from interchanging ω and τ and using assumption (2.4). Hence, if either (2.3) or (2.4) holds, then $|\Re\left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau\right)|$ is either $\Re\left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau\right)$ or $\Re\left(\sigma_0^\tau \Lambda_\tau - \sigma_0^\omega \Lambda_\omega\right)$, and in both cases the lower bound in assertion (2.9) follows from Lemma 2.5. The upper bound in assertion (2.9) follows from the factorization in Lemma 2.4 and the continuity of F_ω and F_τ. □

We also need two known lemmas from previous applications of the factorization method. The first one relates the range of an operator to the norm of its dual or adjoint and the second one shows that the inclusion Ω can be determined from $\mathcal{R}(L)$.

LEMMA 2.7. *Let H_i, $i = 1, 2$, be two Hilbert spaces with norms $\|\cdot\|_i$, X be a third Hilbert space, and $A_i \in \mathcal{L}(X, H_i)$.*

If $\|A_1 x\|_1 \leq \|A_2 x\|_2$ for all $x \in X$, then $\mathcal{R}(A_1^*) \subseteq \mathcal{R}(A_2^*)$ and $\mathcal{R}(A_1') \subseteq \mathcal{R}(A_2')$.

Proof. This follows from the so-called "14th important property of Banach spaces" in Bourbaki [3]; cf. also [10, Lemma 3.4, Cor. 3.5] for an elementary proof for real spaces that holds as well in this complex case. □

LEMMA 2.8. *$\Phi_z|_{\partial B} \in \mathcal{R}(L)$ if and only if $z \in \Omega$.*

Proof. This has been proven by Brühl in [4, Lemma 3.5]. □

We can now prove our main theorems.

Proof of Theorems 2.1 *and* 2.2. Using Lemma 2.7, it follows from Lemma 2.6 that

$$\mathcal{R}\left(\left|\Im\left(\alpha_\omega \Lambda_\omega\right)\right|^{1/2}\right) = \mathcal{R}(L), \quad \text{resp.,} \quad \mathcal{R}\left(\left|\Re\left(\sigma_0^\omega \Lambda_\omega - \sigma_0^\tau \Lambda_\tau\right)\right|^{1/2}\right) = \mathcal{R}(L).$$

Hence, the assertions follow from Lemma 2.8. □

3. Numerical examples.

We tested our method numerically and compared it to the conventional factorization method for static (zero frequency) electrical impedance tomography using exact and inexact reference measurements. B is the two-dimensional unit-disk. The inclusions are two circles centered in $(0.4, 0.2)$ and $(-0.6, 0)$ with radii 0.3 and 0.2. For the complex conductivity at a nonzero frequency ω, we use the values of the first example from Jain et al. in [22]. The background conductivity is $\sigma_0^\omega := 0.3 + 0.1\mathrm{i}$ and inside the inclusions Ω we set this value to $\sigma^\omega|_\Omega := 0.1 + 0.1\mathrm{i}$, i.e., $\sigma_\Omega^\omega := -0.2$. To compare our method with the original factorization method we use $\tau = 0$ as the second frequency and set the imaginary part of the conductivity to zero for that case, i.e., $\sigma_0 := 0.3$ and $\sigma_\Omega := -0.2$. (Consistent with the introduction we omit the index τ for the zero frequency case.) Then

$$\frac{\sigma_\Omega^\omega}{\sigma_0^\omega} = -0.6 + 0.2\mathrm{i} \quad \text{and} \quad \frac{\sigma_\Omega}{\sigma_0} = -\frac{2}{3},$$

so that the assumptions of both Theorems 2.1 and 2.2 are fulfilled. (This also follows from Remark 2.3 as $\tau = 0$ and only the conductivity differs in the inclusions.)

On ∂B we apply

$$\left\{ \frac{1}{\sqrt{\pi}} \sin(n\phi), \, \frac{1}{\sqrt{\pi}} \cos(n\phi) \;\middle|\; n = 1, \ldots, 128 \right\}$$

as input currents, where (r, ϕ) denotes the polar coordinates with respect to the origin.

We use the notation Λ_ω for the NtD operator at the frequency ω and Λ for the corresponding operator at zero frequency (both for the setting with the inclusion Ω). For the original factorization method we also need the reference operator at zero frequency without inclusion Λ_0, i.e., the one corresponding to constant real conductivity $\sigma_0 = 0.3$ throughout B. Accordingly, we denote the corresponding potentials by u, u_ω and u_0.

We calculate these potentials separately using the commercial finite element software Comsol and expand their boundary values in the aforementioned trigonometric basis, which gives us discrete approximations $\tilde{\Lambda}_\omega, \tilde{\Lambda}, \tilde{\Lambda}_0 \in \mathbb{C}^{256\times 256}$. Consistent with our theoretical results in section 2 we describe the applied currents as continuous functions (the so-called continuum model of EIT) and do not study more realistic electrode models in our numerical examples. Also note that we do not directly calculate the difference $\tilde{\Lambda} - \tilde{\Lambda}_0$ as in [13] or $\Im(\sigma_0^\omega \tilde{\Lambda}_\omega)$, resp., $\Re(\sigma_0 \tilde{\Lambda} - \sigma_0^\omega \tilde{\Lambda}_\omega)$ in an analogous manner. Though such a direct calculation of the differences leads to a higher precision in the simulated forward data, we refrained from it in order to be able to simulate independent measurement and shape errors on each of the measurement operators.

The range criteria

$$z \in \Omega \quad \text{if and only if} \quad \Phi_z|_{\partial B} \in \mathcal{R}\left(A^{1/2}\right)$$

with $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ (see Theorem 2.1), $A = |\Re(\sigma_0 \Lambda - \sigma_0^\omega \Lambda_\omega)|$ (see Theorem 2.2) or $A = |\Lambda - \Lambda_0|$ (the conventional factorization method, see Brühl [4, Theorem 3.1]) are implemented as in [13]. For the reader's convenience we repeat the description here. Let

$$A v_k = \lambda_k v_k, \qquad k \in \mathbb{N},$$

be the spectral decomposition of the operator A, which is in all three cases compact, self-adjoint, injective, and positive. $\{v_k\} \subset L^2_\diamond(\partial B)$ is an orthonormal basis of eigenfunctions with eigenvalues $\{\lambda_k\} \subset \mathbb{R}$ (sorted in decreasing order). The Picard criterion yields that

$$\Phi_z|_{\partial B} \in \mathcal{R}(A^{1/2})$$

if and only if

$$f(z) := \frac{1}{\|\Phi_z|_{\partial B}\|^2_{L^2(\partial B)}} \sum_{k=1}^{\infty} \frac{|(\Phi_z|_{\partial B}, v_k)_{L^2(\partial B)}|^2}{\lambda_k} < \infty.$$

Using a singular value decomposition of the discrete approximation $\tilde{A} \in \mathbb{C}^{256\times 256}$ ($\tilde{A} = |\Im(\sigma_0^\omega \tilde{\Lambda}_\omega)|$, $\tilde{A} = |\Re(\sigma_0 \tilde{\Lambda} - \sigma_0^\omega \tilde{\Lambda}_\omega)|$, or $\tilde{A} = |\tilde{\Lambda} - \tilde{\Lambda}_0|$),

$$\tilde{A} \tilde{v}_k = \tilde{\lambda}_k \tilde{u}_k, \qquad \tilde{A}^* \tilde{u}_k = \tilde{\lambda}_k \tilde{v}_k, \qquad k = 1, \dots, 128,$$

with nonnegative $\{\tilde{\lambda}_k\} \subset \mathbb{R}$ (sorted in decreasing order) and orthonormal bases $\{\tilde{u}_k\}, \{\tilde{v}_k\} \subset \mathbb{C}^{256}$, we approximate the function $f(z)$ by

$$\tilde{f}(z) := \sum_{k=1}^{m} \frac{|\tilde{\Phi}_z^* \tilde{v}_k|^2}{\tilde{\lambda}_k} \Big/ \sum_{k=1}^{m} |\tilde{\Phi}_z^* \tilde{v}_k|^2,$$

where $\tilde{\Phi}_z \in \mathbb{C}^{256}$ contains the Fourier coefficients of $\Phi_z|_{\partial B}$, which for the two-dimensional unit circle can be written as (cf., e.g., Brühl [4]),

$$\Phi_z(x) = \frac{1}{\pi}\frac{(z-x)\cdot d}{|z-x|^2} \quad \text{for all } x \in \partial B.$$

m is the number of singular values that are reasonable approximations $\tilde{\lambda}_k \approx \lambda_k$. To estimate m we plot the (normalized) singular values $(\tilde{\lambda}_k)_k$ in a semilogarithmic scale; cf. the left column of Figure 3.1, where this is done for our three different choices for A. Typically, the eigenvalues show an exponential decay that stops rather abruptly due to the presence of errors in our simulated data. In our experiments we manually pick the level where this stop occurs (marked by a dashed line in our eigenvalue plots) and use only the singular values $(\tilde{\lambda}_k)_k$ above this level.

To obtain a numerical criterion telling whether a point z belongs to the unknown inclusion Ω or not, one now has to decide if the infinite sum $f(z)$ attains the value ∞ by using the approximate value $\tilde{f}(z)$, which is always finite. Thus, a threshold $C_\infty > 0$ is needed to distinguish points with *large* values $\tilde{f}(z) \geq C_\infty$ from those with *small* values $\tilde{f}(z) < C_\infty$. A reconstruction of Ω is then obtained by evaluating $\tilde{f}(z)$ on a grid of points $\{z_n\} \subset B$ and saying that all points with $\tilde{f}(z_n) < C_\infty$ belong to the inclusion. Choosing different threshold values C_∞ corresponds to choosing different level contours of $\tilde{f}(z)$ or, equivalently, of a monotone function of $\tilde{f}(z)$.

In our numerical experiments, we plot the indicator function

$$\text{Ind}(z) := \Big(\log\left(1 + \tilde{f}(z)\right)\Big)^{-1} \tag{3.1}$$

on an equidistant grid $\{z_n\} \subset B$, as well as the contour of $\tilde{f}$ that fits best to the true boundary of the inclusion $\partial\Omega$. Note that we choose this optimal contour line in order to compare optimal results for our method with optimal results for the original method. In practice, the choice of a contour line of $\tilde{f}$ has to be done on a heuristic basis or using additional information, e.g., about the size of the inclusion.

3.1. Detecting inclusions without reference measurements. Figure 3.1 shows the reconstructions that we obtain for $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ (top row), $A = |\Re(\sigma_0\Lambda - \sigma_0^\omega \Lambda_\omega)|$ (middle row), and $A = |\Lambda - \Lambda_0|$ (bottom row). The left column shows the (normalized) singular values of A and the trust level marked by a dashed line, the middle column shows the indicator function, and the right column shows the optimal contour line of it (chosen with knowledge of the true inclusions). The true boundary of the inclusions $\partial\Omega$ is plotted with a dashed line in the middle and right columns.

The reconstructions are of similar good quality. For this case of exact simulated data it does not seem to matter whether only measurements at a single nonzero frequency are being used (top row), nonzero frequency measurements are combined with zero frequency measurements (middle row), or zero frequency measurements are combined with reference measurements (bottom row).

In addition to using unperturbed simulated measurements, we also tested the method after adding 0.1% relative noise to the measurement matrix $\tilde{\Lambda}_\omega$, resp., $\tilde{\Lambda}$. More precisely, we generate an error matrix E of the same size as the measurements with uniformly distributed real and imaginary parts of the entries between -1 and 1. E is then scaled to the noise level with respect to its spectral norm and added to the respective measurement operator. (Of course, different errors are added to $\tilde{\Lambda}_\omega$ and $\tilde{\Lambda}$.) We also compare this with the results obtained with the original factorization

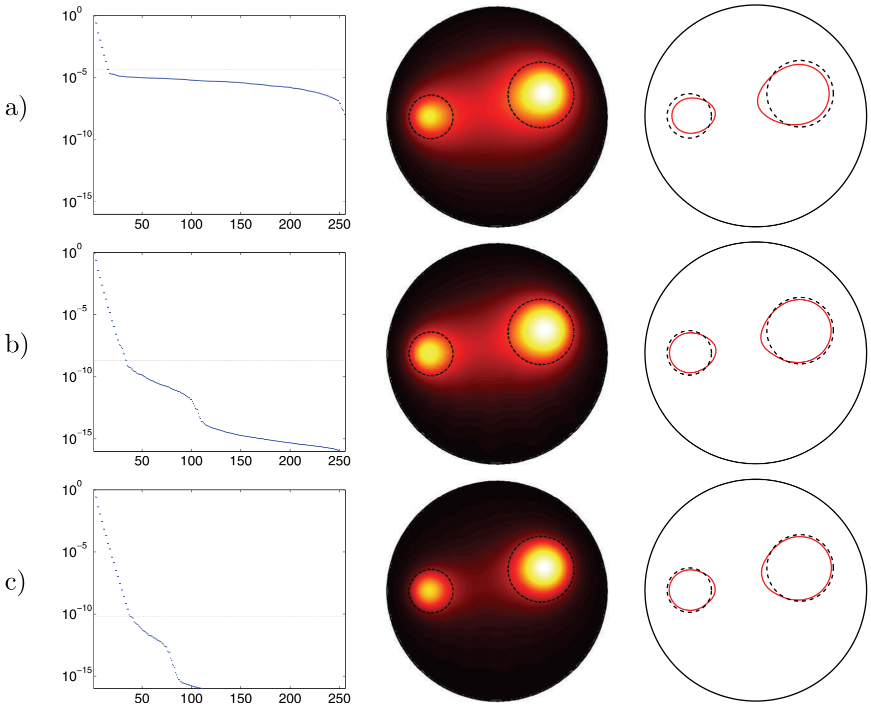


FIG. 3.1. *Numerical results for exact data: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions) for* (a) *single-frequency data,* (b) *frequency-difference data, and* (c) *static data compared with reference data.*

method, i.e., using $\tilde{\Lambda} - \tilde{\Lambda}_0$, where we take the noisy version of $\tilde{\Lambda}$ but not of $\tilde{\Lambda}_0$. The corresponding reconstructions are shown in Figure 3.2, which is organized in the same way as Figure 3.1.

The reconstructions using $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ in the top row and $A = |\Re(\sigma_0 \Lambda - \sigma_0^\omega \Lambda_\omega)|$ in the middle row seem to be more effected by the noise than those using $A = |\Lambda - \Lambda_0|$ in the bottom row. This can be explained by the fact that all three choices for A contain differences of the measurement operators Λ_ω, Λ, or Λ_0, which are much smaller than the measurement operators itself. Thus, the noise on the measurement operators is amplified in these differences. In our example we have that

$$\|\Im(\sigma_0^\omega \tilde{\Lambda}_\omega)\| \approx 0.06 \, \|\sigma_0^\omega \tilde{\Lambda}_\omega\| \, ,$$

$$\|\Re(\sigma_0 \tilde{\Lambda} - \sigma_0^\omega \tilde{\Lambda}_\omega)\| \approx 0.03 \, \|\sigma_0 \tilde{\Lambda}\| \approx 0.03 \, \|\sigma_0 \tilde{\Lambda}\| \, ,$$

$$\|\tilde{\Lambda} - \tilde{\Lambda}_0\| \approx 0.14 \, \|\tilde{\Lambda}\| \approx 0.16 \, \|\tilde{\Lambda}_0\| \, ,$$

so that the noise amplification is highest in the middle row and lowest in the bottom row, which fits well to the different quality of the reconstruction. Note that this amount of noise amplification depends on the inclusions' contrast to the background conductivity and, though we have not thoroughly investigated this question, different numerical examples did not indicate that one choice of A is generally more robust to noise than another.

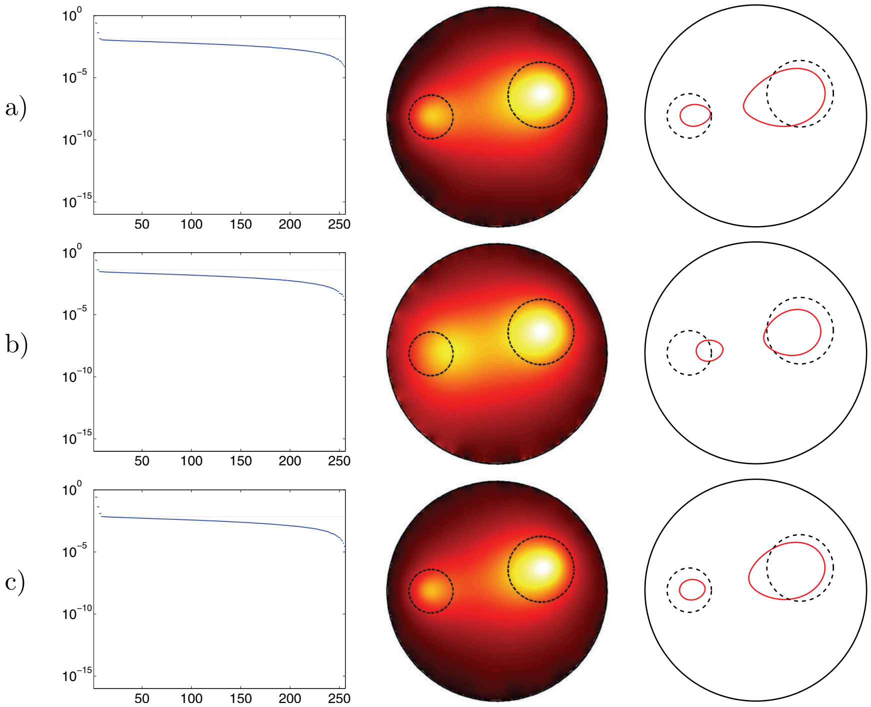


FIG. 3.2. *Numerical reconstructions for data containing* 0.1% *relative noise: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions) for* (a) *single-frequency data,* (b) *frequency-difference data, and* (c) *static data compared with reference data.*

3.2. Sensitivity to body shape errors. We also compared our method to the conventional factorization method in a setting where there are some boundary geometry errors between the computational domain of the forward model and that of the true body. To that end we replaced B by an ellipse with halfaxes $1+\delta$ and $1/(1+\delta)$ in the calculation of the forward data. Everything else remained unchanged to simulate the case where this ellipse is wrongly assumed to be the unit circle.

We used $\delta = 5\%$ and show in the first row of Figure 3.3 the reconstruction obtained from $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ and in the second row those obtained with $A = |\Re(\sigma_0 \Lambda - \sigma_0^\omega \Lambda_\omega)|$. As we explained in the introduction, for the conventional factorization method, using $A = |\Lambda - \Lambda_0|$, the effect of body shape errors strongly depends on whether the reference operator Λ_0 is experimentally obtainable or numerically simulated (resp., in easy cases, calculated analytically). In the first case, Λ_0 correctly corresponds to measurements at an ellipse (which the body really is), while in the latter case, it corresponds to a circle (which we wrongly assume the body to be). The resulting reconstructions are shown in the third and forth row of Figure 3.3.

As we expected, systematic body shape errors have a greater effect on the reconstructions when the measurement operators belong to a different geometry, as is the case for the conventional factorization method with simulated reference data, shown in the bottom row. However, also using only measurements at a single, nonzero frequency, i.e., $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ in the top row, seems to be similarly effected. If two different kinds of measurements are taken at the same body, the reconstructions improve. It does not seem to matter much whether these two are measurements at a

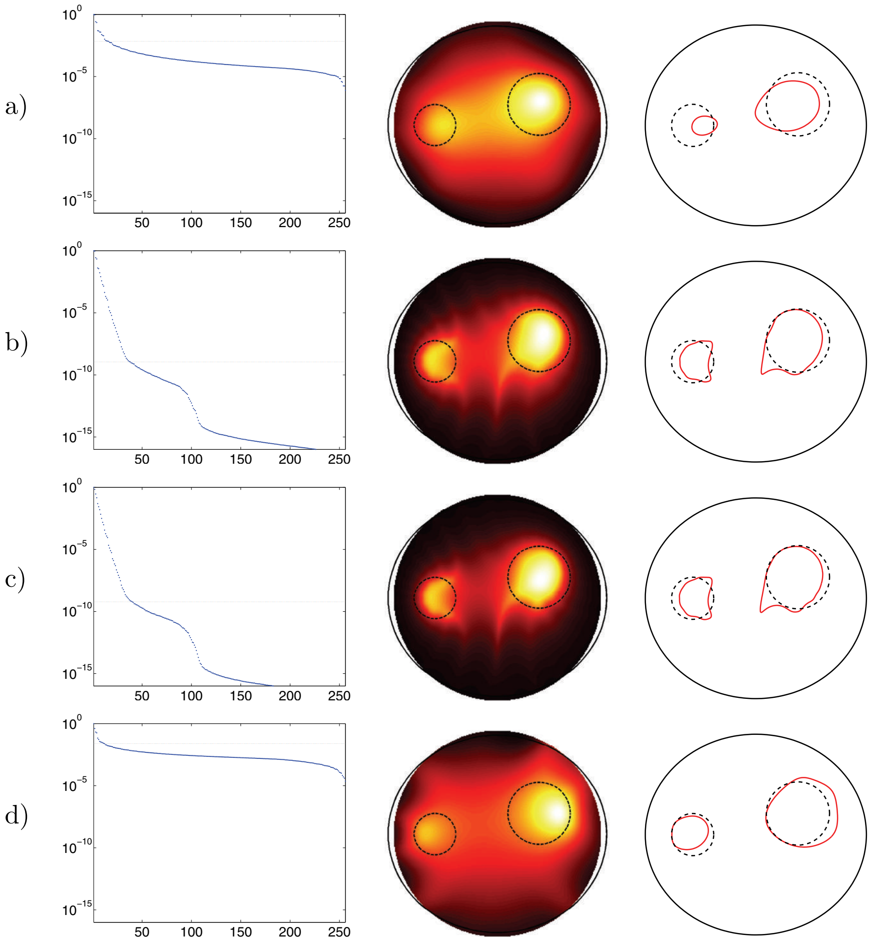


Fig. 3.3. *Numerical reconstructions for data containing $\delta = 5\%$ body shape errors: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions) for* (a) *single-frequency data,* (b) *frequency-difference data,* (c) *static data compared with reference data for the correct body shape, and* (d) *static data compared with reference data for the assumed (incorrect) body shape.*

nonzero and at zero frequency ($A = |\Re(\sigma_0\Lambda - \sigma_0^\omega\Lambda_\omega)|$ in the second row) or zero frequency measurements with and without inclusion ($A = |\Lambda - \Lambda_0|$ in the third row).

Figure 3.4, which is organized in the same way as Figure 3.3, shows the reconstructions that we obtained for an ellipse with halfaxes $1+\delta$ and $1/(1+\delta)$ for $\delta = 10\%$. Even for this rather large amount of errors in the estimated body shape the reconstructions are quite reasonable if frequency-difference measurements (second row) or the conventional factorization method is used with reference measurements belonging to the exact (elliptical) body shape (third row). If measurements at only a single, nonzero frequency are used (top row), the reconstruction gets highly blurred. The conventional method with reference measurements that are simulated for the assumed (incorrect) body shape (bottom row) performs even worse and leads to strong artifacts.

The results numerically verify that our new variant of the factorization method with frequency-difference data is much more robust against body shape errors than

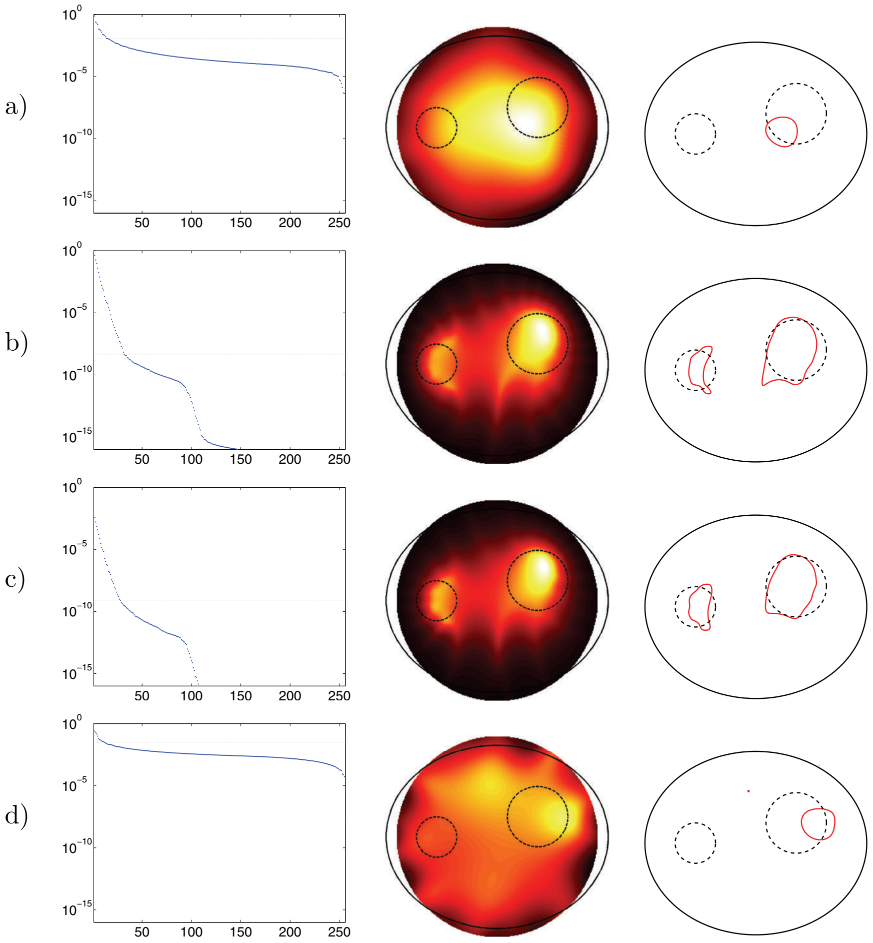


Fig. 3.4. *Numerical reconstructions for data containing $\delta = 10\%$ body shape errors: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions) for* (a) *single-frequency data,* (b) *frequency-difference data,* (c) *static data compared with reference data for the correct body shape, and* (d) *static data compared with reference data for the assumed (incorrect) body shape.*

using the conventional method with simulated (or analytically calculated) reference data. Actually, the reconstructions using frequency-difference data seem to be of equally good quality and robustness as those that one would obtain with correct reference measurements (which are usually not available in practice).

3.3. Unknown background conductivity. Though our new variant of the factorization method works without reference measurements, it still requires the knowledge of the constant conductivity value of the background. We now describe a heuristic approach with which the algorithm can also be applied to an unknown background conductivity.

Roughly speaking, we expect that fast spatial variations in the applied currents on ∂B lead to higher electric currents close to ∂B, while the electric effect of slowly spatially varying currents penetrates deeper into B (see [12] for a detailed study of how to create potentials with localized electrical energy). We also expect that the eigenfunctions of our compact measurement operators Λ^ω, resp., Λ, are functions containing increasingly high oscillations. Thus, the eigenvalues will mostly depend on the conductivity close to ∂B, i.e., the background conductivity, σ_0^ω, resp., σ_0.

The multiplication of Λ_ω with the background value σ_0^ω can be regarded as a simple change of units that sets this background conductivity value to the real value 1. Thus, we expect the eigenvalues of $\sigma_0^\omega \Lambda_\omega$ to mostly lie close to the real axis, and so the eigenvalues of Λ_ω will lie close to a straight line through the origin whose angle with the real axis is minus the phase of σ_0^ω.

To determine the range of $|\Im(\sigma_0^\omega \Lambda_\omega)|^{1/2}$, it suffices to estimate this phase of σ_0^ω. We do this by choosing the median α of the set

$$\left\{ -\frac{\Im(\tilde{\lambda}_j^\omega)}{\Re(\tilde{\lambda}_j^\omega)} \ : \ j = 1, 2, \dots \right\}$$

using all available eigenvalues $\tilde{\lambda}_j^\omega$ of $\tilde{\Lambda}_\omega$. Instead of $A = |\Im(\sigma_0^\omega \Lambda_\omega)|$ we then use $A = |\Im((1 + \alpha\mathrm{i})\Lambda_\omega)|$ in our algorithm.

To estimate the range of $|\Re(\sigma_0 \Lambda - \sigma_0^\omega \Lambda_\omega)|^{1/2}$ we proceed analogously and estimate the quotient of σ_0^ω and σ_0 by the median β of the set

$$\left\{ \frac{\tilde{\lambda}_j}{\tilde{\lambda}_j^\omega} \ : \ j = 1, 2, \dots \right\}$$

using all available eigenvalues $\tilde{\lambda}_j^\omega$ of $\tilde{\Lambda}_\omega$ and $\tilde{\lambda}_j$ of $\tilde{\Lambda}$. Then we use $A = |\Re(\Lambda - \beta\Lambda_\omega)|$ in our algorithm.

Figure 3.5 shows the reconstructions that we obtained with this approach for our numerical example. The columns are organized as in Figures 3.1–3.4. The first two rows show the reconstructions obtained with $A = |\Im((1 + \alpha\mathrm{i})\Lambda_\omega)|$ and $A = |\Re(\Lambda - \beta\Lambda_\omega)|$ and the last two rows show the according reconstructions after adding 0.1% of relative noise as in subsection 3.1. The reconstructions show almost no visual difference to those obtained with the exact background conductivity values in subsection 3.1.

Though this is not covered by the theory presented in section 2, it seems plausible that the above approach can also be applied to cases where the unknown background is slightly inhomogeneous. In our final example we test this numerically for the factorization method with frequency-difference data. We multiply the complex background conductivity used in the previous examples with the slightly oscillating function

$$1 + 0.05\cos(4\pi x)\sin(8\pi y).$$

In order to retain the conductivity jump only in the real part, we also multiply the imaginary part of the conductivity inside the inclusions with this function. Figure 3.6 shows the resulting real part (left picture) and the imaginary part (right picture) of σ^ω. As in the previous examples, we assume that the static conductivity σ has the same real part as σ^ω and that it has zero imaginary part.

We also added $\delta = 5\%$ body shape error as in subsection 3.2 and 0.05% relative noise as in subsection 3.1. Figure 3.7 shows the reconstruction that we obtain from using frequency-difference data $A = |\Re(\Lambda - \beta\Lambda_\omega)|$, where the constant $\beta \in \mathbb{C}$ is estimated from the data as explained above.

The reconstruction are comparable to those obtained with knowledge of the exact body shape, known constant background, and 0.1% relative noise in subsection 3.1. This suggests that the method can indeed be applied to the practically relevant case of an unknown inhomogeneous background that is only known to be “almost constant.”

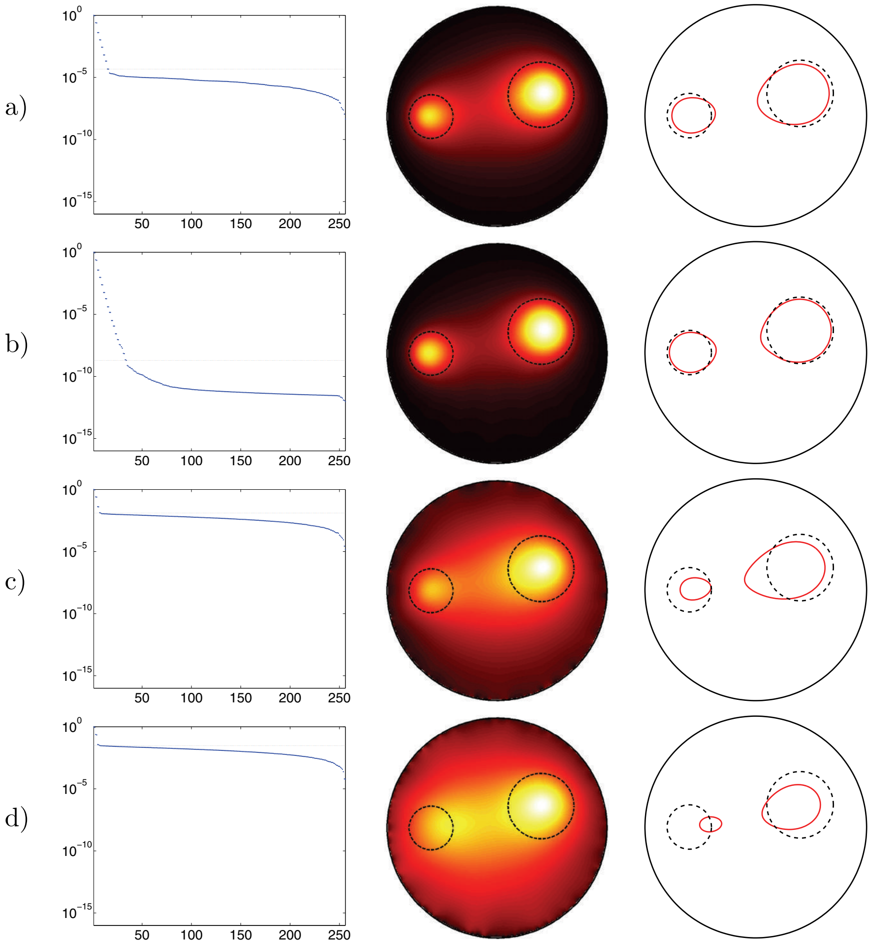


FIG. 3.5. *Numerical reconstructions for an unknown, but constant, background conductivity: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions) for* (a) *exact single-frequency data,* (b) *exact frequency-difference data,* (c) *single-frequency data containing* 0.1% *relative noise, and* (d) *frequency-difference data containing* 0.1% *relative noise.*

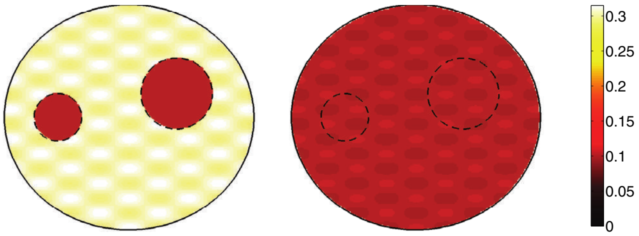


FIG. 3.6. *Real and imaginary part of the conductivity describing inclusions in a slightly inhomogeneous background.*

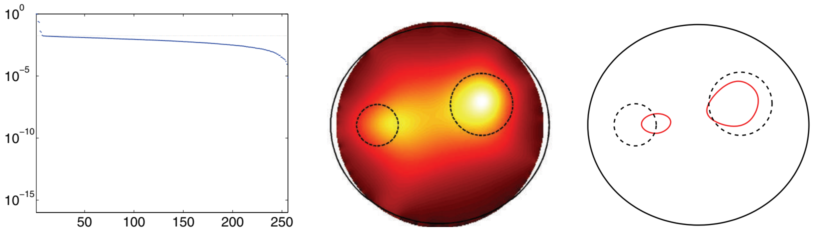


Fig. 3.7. *Numerical reconstructions for an unknown, slightly inhomogeneous background conductivity using frequency-difference data containing $\delta = 5\%$ body shape errors and 0.05% relative noise: singular values (left column), indicator function (middle column), and its optimal contour (right column, chosen with knowledge of the true inclusions).*

4. Conclusions.

We have developed a new variant of the factorization method that can be used on single-frequency and on frequency-difference measurements in electrical impedance tomography and that does not require reference measurements at an inclusion-free body, which are usually not available in practice. Our new variant with single-frequency measurements delivers comparable results to using the conventional method with simulated reference data and thus eliminates one of the main computational efforts in applying the method. An even greater advantage is achieved by using frequency-difference measurements. Not only do we save the computational effort of simulating reference measurements, but our new results show the same performance in the presence of body shape errors that one would otherwise obtain from reference measurements at the correct (unknown) body shape. This greatly improves the stability of the method with respect to such unavoidable systematic errors, so that the application of the method in frequency-difference EIT systems seems very promising.

REFERENCES

[1] H. Ammari, R. Griesmaier, and M. Hanke, *Identification of small inhomogeneities: Asymptotic factorization*, Math. Comp., 76 (2007), pp. 1425–1448.

[2] H. Ammari and J. K. Seo, *An accurate formula for the reconstruction of conductivity inhomogeneities*, Adv. in Appl. Math., 30 (2003), pp. 679–705.

[3] N. Bourbaki, *Elements of Mathematics, Topological Vector Spaces*, Chapters 1–5, Springer-Verlag, Berlin, 2003.

[4] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[5] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[6] M. Brühl, M. Hanke, and M. S. Vogelius, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[7] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology—Volume 2: Functional and Variational Methods*, Springer-Verlag, Berlin, 2000.

[8] H. Eckel and R. Kress, *Nonlinear integral equations for the inverse electrical impedance problem*, Inverse Problems, 23 (2007), pp. 475–491.

[9] K. Erhard and R. Potthast, *The point source method for reconstructing an inclusion from boundary measurements in electrical impedance tomography and acoustic scattering*, Inverse Problems, 19 (2003), pp. 1139–1157.

[10] F. Frühauf, B. Gebauer, and O. Scherzer, *Detecting interfaces in a parabolic-elliptic problem from surface measurements*, SIAM J. Numer. Anal., 45 (2007), pp. 810–836.

[11] B. Gebauer, *The factorization method for real elliptic problems*, Z. Anal. Anwend., 25 (2006), pp. 81–102.

[12] B. Gebauer, *Localized potentials in electrical impedance tomography*, Inverse Probl. Imaging,

2 (2008), pp. 251–269.

[13] B. Gebauer and N. Hyvönen, *Factorization method and irregular inclusions in electrical impedance tomography*, Inverse Problems, 23 (2007), pp. 2159–2170.

[14] M. Hanke and M. Brühl, *Recent progress in electrical impedance tomography*, Inverse Problems, 19 (2003), pp. S65–S90.

[15] M. Hanke, N. Hyvönen, and S. Reusswig, *Convex source support and its application to electric impedance tomography*, SIAM J. Imaging Sci., 1 (2008), pp. 364–378.

[16] M. Hanke and B. Schappel, *The factorization method for electrical impedance tomography in the half-space*, SIAM J. Appl. Math., 68 (2008), pp. 907–924.

[17] F. Hettlich and W. Rundell, *The determination of a discontinuity in a conductivity from a single boundary measurement*, Inverse Problems, 14 (1998), pp. 67–82.

[18] N. Hyvönen, *Complete electrode model of electrical impedance tomography: Approximation properties and characterization of inclusions*, SIAM J. Appl. Math., 64 (2004), pp. 902–931.

[19] N. Hyvönen, H. Hakula, and S. Pursiainen, *Numerical implementation of the factorization method within the complete electrode model of electrical impedance tomography*, Inverse Probl. Imaging, 1 (2007), pp. 299–317.

[20] T. Ide, H. Isozaki, S. Nakata, S. Siltanen, and G. Uhlmann, *Probing for electrical inclusions with complex spherical waves*, Comm. Pure Appl. Math., 60 (2007), pp. 1415–1442.

[21] K. Ito, K. Kunisch, and Z. Li, *Level-set function approach to an inverse interface problem*, Inverse Problems, 17 (2001), pp. 1225–1242.

[22] H. Jain, D. Isaacson, P. M. Edic, and J. C. Newell, *Electrical impedance tomography of complex conductivity distributions with noncircular boundary*, IEEE Trans. Biomedical Engineering, 44 (1997), pp. 1051–1060.

[23] H. Kang, J. K. Seo, and D. Sheen, *The inverse conductivity problem with one measurement: Stability and estimation of size*, SIAM J. Math. Anal., 28 (1997), pp. 1389–1405.

[24] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[25] A. Kirsch, *The factorization method for a class of inverse elliptic problems*, Math. Nachr., 278 (2005), pp. 258–277.

[26] A. Kirsch and N. Grinberg, *The Factorization Method for Inverse Problems*, Oxford Lecture Ser. Math. Appl. 36, Oxford University Press, Oxford, 2008.

[27] O. Kwon, J. K. Seo, and J. R. Yoon, *A real-time algorithm for the location search of discontinuous conductivities with one measurement*, Comm. Pure Appl. Math., 55 (2002), pp. 1–29.

[28] O. Kwon, J. R. Yoon, J. K. Seo, E. J. Woo, and Y. G. Cho, *Estimation of anomaly location and size using electrical impedance tomography*, IEEE Trans. Biomed. Eng., 50 (2003), pp. 89–96.

[29] A. Lechleiter, N. Hyvönen, and H. Hakula, *The factorization method applied to the complete electrode model of impedance tomography*, SIAM J. Appl. Math., 68 (2008), pp. 1097–1121.

[30] A. I. Nachman, L. Päivärinta, and A. Teirilä, *On imaging obstacles inside inhomogeneous media*, J. Funct. Anal., 252 (2007), pp. 490–516.

[31] T. I. Oh, J. Lee, J. K. Seo, S. W. Kim, and E. J. Woo, *Feasibility of breast cancer lesion detection using a multi-frequency trans-admittance scanner (tas) with* 10 *hz to* 500 *khz bandwidth*, Physiol. Meas., 28 (2007), pp. S71–S84.

[32] J. K. Seo, J. Lee, H. Zribi, S. W. Kim, and E. J. Woo, *Frequency-difference electrical impedance tomography (fdEIT): Algorithm development and feasibility study*, Physiol. Meas., 29 (2008), pp. 929–944.